entropy

MDPI

Article

Measuring Integrated Information: Comparison of Candidate Measures in Theory and Simulation

Pedro A.M. Mediano [1,*], Anil K. Seth [2] and Adam B. Barrett [2]

[1] Department of Computing, Imperial College, London SW7 2RH, UK

[2] Sackler Centre for Consciousness Science and Department of Informatics, University of Sussex, Brighton BN1 9RH, UK; a.k.seth@sussex.ac.uk (A.K.S.); adam.barrett@sussex.ac.uk (A.B.B.)

* Correspondence: pmediano@imperial.ac.uk; Tel.: +44-20-759-48445

Received: 11 September 2018; Accepted: 18 December 2018; Published: 25 December 2018

Abstract: Integrated Information Theory (IIT) is a prominent theory of consciousness that has at its centre measures that quantify the extent to which a system generates more information than the sum of its parts. While several candidate measures of integrated information ("Φ") now exist, little is known about how they compare, especially in terms of their behaviour on non-trivial network models. In this article, we provide clear and intuitive descriptions of six distinct candidate measures. We then explore the properties of each of these measures in simulation on networks consisting of eight interacting nodes, animated with Gaussian linear autoregressive dynamics. We find a striking diversity in the behaviour of these measures—no two measures show consistent agreement across all analyses. A subset of the measures appears to reflect some form of dynamical complexity, in the sense of simultaneous segregation and integration between system components. Our results help guide the operationalisation of IIT and advance the development of measures of integrated information and dynamical complexity that may have more general applicability.

Keywords: integrated information theory; computational neuroscience; complexity; consciousness

1. Introduction

Measures of integrated information seek to quantify the extent to which a whole system generates more information than the sum of its parts as it transitions between states. In biological systems, integrated information could underpin cognitive and behavioural flexibility, and even consciousness. More generally, integrated information measures have the potential to capture the dynamical complexity of any many body system, and hence to aid with understanding and characterising complex systems [1]. Since the concept of integrated information can be operationalised in many different ways, a whole range of distinct integrated information measures have come into being in the literature [2–5]. Several of them are beginning to see application to empirical data [6], or to large-scale simulations [7,8], yet a systematic comparison of the behaviour of the various measures on non-trivial network models has not previously been performed.

This paper has two goals: first, to provide a unified source of explanation of the principles and practicalities of a class of prominent candidate measures of integrated information; second, to examine the behaviour of candidate measures on non-trivial network models, in order to shed light on their comparative practical utility.

The class of integrated information measures we consider could be called *dynamical*, or *empirical*, integrated information measures. Following Barrett and Seth [2], they quantify the information that the current state contains about a past state (for the information integrated over time window τ, the past state to be considered is that at time τ from the present), measured using the empirical, or spontaneous, distribution for *a priori* uncertainty about the past state. These measures are well-defined for any

stochastic system (with a well-defined Lebesgue measure across the states) and have the advantage that they can be estimated for real data using empirical distributions if stationarity can be assumed. By contrast, the integrated information measures introduced in especially more recent versions of the Integrated Information Theory of consciousness (IIT) are applicable only to discrete Markovian systems [9]. These measures, which could be called *causal* integrated information measures, compute the information generated when the system transitions to one particular state out of a repertoire of possible states; the *a priori* distribution for the past state is taken to be maximum entropy, so that the measures reflect every possible thing that the system could do.

Related to integrated information is the notion of dynamical complexity, which has been variously defined [10,11]. Sometimes dynamical complexity is considered as the whole being somehow greater than the sum of parts, and sometimes as a system showing a balance between two competing tendencies, namely

- **integration**, i.e., the system behaves as one; and
- **segregation**, i.e., the parts of the system behave independently.

The notion of dynamical complexity has further been described as a balance between order and disorder, or between chaos and synchrony, and has been related to criticality and metastability [7]. Similarly, it has been argued that a necessary feature of complexity measures is to peak in systems that exhibit a mixture between low and high correlation [11]. Many quantitative measures of dynamical complexity have been proposed, but a theoretically-principled, one-size-fits-all measure remains elusive. In exploring the behaviour of each of the integrated information measures, we consider to what extent they reflect integration, segregation, and a balanced degree of correlation.

In short, measures of integrated information and dynamical complexity share the notion of tracking the extent to which the whole is more than the sum of the parts. While early versions of IIT explicitly linked integrated information with dynamical complexity [10], later versions have focused on a "causal" definition of integrated information which has more to do with the "irreducibility of mechanisms" than the co-existence of integration and segregation in a system's dynamics [9]. Here, we focus on the earlier, dynamical/empirical conceptions of integrated information because (i) they are more readily applicable to empirical time-series; (ii) they remain conceptually powerful in theories of consciousness, and (iii) they promise general applicability to many other questions in neuroscience and beyond, in which part-whole relations are of interest.

We consider five distinct and prominent measures of dynamical integrated information: whole-minus-sum integrated information Φ [12]; integrated stochastic interaction $\tilde{\Phi}$ [2]; integrated synergy ψ [3]; decoder-based integrated information Φ^* [5]; and geometric integrated information Φ_G [4]. We also consider, for comparison, the measure causal density (CD) [13,14], which can be considered as the sum of independent information transfers in the system (without reference to a minimum information partition). This measure has previously been discussed in conjunction with integrated information and dynamical complexity measures [13,15]. For a comparison of related complexity measures, see Reference [16].

All of the measures have the potential to behave in ways which are not obvious a priori, and in a manner difficult to express analytically. While some simulations of some of the measures (Φ, $\tilde{\Phi}$, CD) on networks have been performed [2,13], and some analytical understanding has been achieved for Φ and $\tilde{\Phi}$ [5,17], other measures (Φ^*, Φ_G) have not previously been computed on any model consisting of more than two components. This paper provides a comparison of the full suite of measures on non-trivial network models. We consider eight-node networks with a range of different architectures, animated with basic noisy vector autoregressive dynamics. We examine how each measure is affected by network topology, coupling strength and noise correlation, as well as its relation with global correlation (a simple dynamical control). Based on these comparisons, we discuss the extent to which each measure captures the co-existence of integration and segregation central to the concept of dynamical complexity.

After covering the necessary preliminaries, we set out the intuition behind the measures, and summarise the mathematics behind the definition of each measure. Next, we present the

simulations and their results, and conclude with a discussion of the implications for IIT. In the Appendices, we derive new formulae for computing the decoder-based integrated information Φ^* for Gaussian systems, correcting the previous formulae in Reference [5], and present further derivations of mathematical properties of the measures.

2. Methods

2.1. Notation, Convention and Preliminaries

In this section, we review the fundamental concepts needed to define and discuss the candidate measures of integrated information. For a more comprehensive introduction, see Reference [18]. In general, we will denote random variables with uppercase letters (e.g., X, Y) and particular instantiations with the corresponding lowercase letters (e.g., x, y). Variables can be either continuous or discrete, and we assume that continuous variables can take any value in $\mathbb{R}^n$ and that a discrete variable X can take any value in the finite set Ω_X. Whenever there is a sum involving a discrete variable X, we assume the sum runs for all possible values of X (i.e., the whole Ω_X). A partition $\mathcal{P} = \{M^1, M^2, \ldots, M^r\}$ divides the elements of system X into r non-overlapping, non-empty sub-systems (or parts), such that $X = M^1 \bigcup M^2 \bigcup \cdots \bigcup M^r$ and $M^i \bigcap M^j = \emptyset$, for any i, j. We denote each variable in X as X^i, and the total number of variables in X as n. When dealing with time series, time will be indexed with a subscript, e.g., X_t.

Entropy H quantifies the uncertainty associated with random variable X—i.e., the higher $H(X)$ the harder it is to make predictions about X—and is defined as

$$H(X) =: -\sum_x p(x) \log p(x). \tag{1}$$

In many scenarios, a discrete set of states is insufficient to represent a process or time series. This is the case, for example, with brain recordings, which come in real-valued time series and with no a priori discretisation scheme. In these cases, using a continuous variable $X \in \mathbb{R}$, we can similarly define the *differential entropy*,

$$H[p] =: -\int p(x) \log p(x) dx. \tag{2}$$

However, differential entropy is not as interpretable and well-behaved as its discrete-variable counterpart. For example, differential entropy is not invariant to rescaling or other transformations on X. Moreover, it is only defined if X has a density with respect to the Lebesgue measure dx; this assumption will be upheld throughout this paper. We can also define the *conditional* and *joint* entropies as

$$\begin{aligned} H(X \mid Y) &=: \sum_y p(y) H(X \mid Y = y) \\ &= -\sum_y p(y) \sum_x p(x \mid y) \log p(x \mid y), \end{aligned} \tag{3}$$

$$H(X, Y) =: -\sum_{x,y} p(x, y) \log p(x, y), \tag{4}$$

respectively. Conditional and joint entropies can be analogously defined for continuous variables by appropriately replacing sums with integrals.

The Kullback–Leibler (KL) divergence quantifies the dissimilarity between two probability distributions p and q:

$$D_{KL}(p \| q) =: \sum_x p(x) \log \frac{p(x)}{q(x)}. \tag{5}$$

The KL divergence represents a notion of (non-symmetric) distance between two probability distributions. It plays an important role in information geometry, which deals with the geometric structure of manifolds of probability distributions.

Finally, mutual information I quantifies the interdependence between two random variables X and Y. It is the KL divergence between the full joint distribution and the product of marginals, but it can also be expressed as the average reduction in uncertainty about X when Y is given:

$$\begin{aligned} I(X;Y) &=: D_{KL}\left(p(X,Y) \parallel p(X)p(Y)\right) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y). \end{aligned} \tag{6}$$

Mutual information is symmetric in the two arguments X and Y. We make use of the following properties of mutual information:

1. $I(X;Y) = I(Y;X)$,
2. $I(X;Y) \geq 0$, and
3. $I(f(X);g(Y)) = I(X;Y)$ for any injective functions f, g.

We highlight one implication of property 3: I is upper-bounded by the entropy of both X and Y. This means that the entropy $H(X)$ of a random variable X is the maximum amount of information X can have about any other variable Y (or another variable Y can have about X).

Mutual information is defined analogously for continuous variables and, unlike differential entropy, it retains its interpretability in the continuous case [19]. Furthermore, one can track how much information a system preserves during its temporal evolution by computing the time-delayed mutual information (TDMI) $I(X_t; X_{t-\tau})$.

Next, we introduce notation and several useful identities to handle Gaussian variables. Given an n-dimensional real-valued system X, we denote its covariance matrix as $\Sigma(X)_{ij} =: \mathrm{cov}(X^i, X^j)$. Similarly, cross-covariance matrices are denoted as $\Sigma(X,Y)_{ij} =: \mathrm{cov}(X^i, Y^j)$. We will make use of the conditional (or partial) covariance formula,

$$\Sigma(X|Y) =: \Sigma(X) - \Sigma(X,Y)\Sigma(Y)^{-1}\Sigma(Y,X). \tag{7}$$

For Gaussian variables,

$$H(X) = \frac{1}{2}\log(\det\Sigma(X)) + \frac{1}{2}n\log(2\pi e)\,, \tag{8}$$

$$H(X|Y=y) = \frac{1}{2}\log(\det\Sigma(X|Y)) + \frac{1}{2}n\log(2\pi e)\,,\ \ \forall y\,, \tag{9}$$

$$I(X;Y) = \frac{1}{2}\log\left(\frac{\det\Sigma(X)}{\det\Sigma(X|Y)}\right). \tag{10}$$

All systems we deal with in this article are stationary and ergodic, so throughout the paper $\Sigma(X_t) = \Sigma(X_{t-\tau})$ for any τ.

2.2. Integrated Information Measures

2.2.1. Overview

In this section, we review the theoretical underpinnings and practical considerations of several proposed measures of integrated information, and in particular how they relate to intuitions about segregation, integration and complexity. These measures are:

- Whole-minus-sum integrated information, Φ;
- Integrated stochastic interaction, $\tilde{\Phi}$;
- Integrated synergy, ψ;

- Decoder-based integrated information, Φ^*;
- Geometric integrated information, Φ_G; and
- Causal density, CD.

All of these measures (besides CD) have been inspired by the measure proposed by Balduzzi and Tononi in [12], which we call Φ_{2008}. Φ_{2008} was based on the information the current state contains about a hypothetical maximum entropy past state. In practice, this results in measures that are applicable only to discrete Markovian systems [2]. For broader applicability, it is more practical to build measures based on the ongoing spontaneous information dynamics—that is, based on $p(X_t, X_{t-\tau})$ without applying a perturbation to the system. Measures are then well-defined for any stochastic system (with a well-defined Lebesgue measure across the states), and can be estimated for real data using empirical distributions if stationarity can be assumed. All of the measures we consider in this paper are based on a system's spontaneous information dynamics.

Table 1 contains a brief description of each measure and a reference to the original publication that introduced it [20]. We refer the reader to the original publications for more detailed descriptions of each measure. Table 2 contains a summary of properties of the measures considered, proven for the case in which the system is ergodic and stationary, and the spontaneous distribution is used.

Table 1. Integrated information measures considered and original references.

Measure	Description	Reference
Φ	Information lost after splitting the system	[12]
$\tilde{\Phi}$	Uncertainty gained after splitting the system	[2]
ψ	Synergistic predictive information between parts of the system	[3]
Φ^*	Past state decoding accuracy lost after splitting the system	[5]
Φ_G	Information-geometric distance to system with disconnected parts	[4]
CD	Average pairwise directed information flow	[13]

Table 2. Overview of properties of integrated information measures, proofs in Appendix C.

	Φ	$\tilde{\Phi}$	ψ	Φ^*	Φ_G	CD
Time-symmetric	✓	✓	×	?	✓	×
Non-negative	×	✓	✓	✓	✓	✓
Invariant to variable rescaling	✓	×	✓	✓	✓	✓
Upper-bounded by time-delayed mutual information	✓	×	✓	✓	✓	✓
Known estimators for arbitrary real-valued systems	✓	✓	×	×	×	✓
Closed-form expression in discrete and Gaussian systems	✓	✓	✓	×	×	✓

2.2.2. Minimum Information Partition

Key to all measures of integrated information is the notion of splitting or partitioning the system to quantify the effect of such split on the system as a whole. In that spirit, integrated information measures are defined through some measure of *effective information*, which operationalises the concept of "information *beyond* a partition" $\mathcal{P}$. This typically involves splitting the system according to $\mathcal{P}$ and computing some form of information loss, via (for example) mutual information (Φ), conditional entropy ($\tilde{\Phi}$), or decoding accuracy (Φ^*) (see Table 1). Integrated information is then the effective information with respect to the partition that identifies the "weakest link" in the system, i.e., the partition for which the parts are least integrated. Formally, integrated information is the effective information beyond the *minimum information partition* (MIP), which, given an effective information measure $f[X; \tau, \mathcal{P}]$, is defined as

$$\mathcal{P}_{\text{MIP}} = \arg_{\mathcal{P}} \min \frac{f[X; \tau, \mathcal{P}]}{K(\mathcal{P})}, \tag{11}$$

where $K(\mathcal{P})$ is a normalisation coefficient. In other words, the MIP is the partition across which the (normalised) effective information is minimum, and integrated information is the (unnormalised) effective information beyond the MIP. The purpose of the normalisation coefficient is to avoid biasing the minimisation towards unbalanced bipartitions (recall that the extent of information sharing between parts is bounded by the entropy of the smaller part). Balduzzi and Tononi [12] suggest the form

$$K(\mathcal{P}) = (r-1) \min_k H(M_t^k). \tag{12}$$

However, not all contributions to IIT have followed Balduzzi and Tononi's treatment of the MIP. Of the measures listed above, Φ and $\tilde{\Phi}$ share this partition scheme, ψ defines the MIP through an *unnormalised* effective information, and Φ^*, Φ_G and CD are defined via the atomic partition without any reference to the MIP. These differences are a confounding factor when it comes to comparing measures—it becomes difficult to ascertain whether differences in behaviour of various measures are due to their definitions of effective information, to their normalisation factor (or lack thereof), or to their partition schemes. We return to this topic in the Discussion section below.

In the following, we present all measures as they were introduced in their original papers (see Table 1), although it is trivial to combine different effective information measures with different partition optimisation schemes. However, all results presented here are calculated by minimising each unnormalised effective information measure over even-sized bipartitions—i.e., bipartitions in which both parts have the same number of components. This is to avoid conflating the effect of the partition scan method with the effect of the integrated information measure itself.

2.2.3. Whole-Minus-Sum Integrated Information Φ

We next turn to the different measures of integrated information. As highlighted above, a primary difference among them is how they define the effective information beyond a given partition. Since most measures were inspired by Balduzzi and Tononi's Φ_{2008}, we start there.

For Φ_{2008}, the effective information φ_{2008} is written as (following notation from [21]) the KL divergence between $p_c(X_0|X_1 = x)$ and $\Pi_k p_c(M_0^k|M_1^k = m^k)$, where $p_c(X_0|X_1 = x)$ (and analogously $p_c(M_0^k|M_1^k = m^k)$) is the conditional distribution for X_0 given $X_1 = x$ under the perturbation at time 0 into all states with equal probability—i.e., given that the joint distribution is given by $p_{ce}(X_0, X_1) =: p(X_1|X_0)p_u(X_0)$, where p_u is the uniform (maximum entropy) distribution [22].

Averaging φ_{2008} over all states x, the result can be expressed as either

$$I(X_0; X_1) - \sum_{k=1}^{r} I(M_0^k; M_1^k), \tag{13}$$

or

$$-H(X_0|X_1) + \sum_{k=1}^{r} H(M_0^k|M_1^k). \tag{14}$$

These two expressions are equivalent under the uniform perturbation, since they differ only by a factor that vanishes if $p(X_0)$ is the uniform distribution. However, they are *not* equivalent if the spontaneous distribution of the system is used instead—i.e., if $p(X_{t-\tau}, X_t)$ is used instead of $p_{ce}(X_0, X_1)$. This means that, for application to spontaneous dynamics (i.e., without perturbation), we have two alternatives that give rise to two measures that are both equally valid analogs of Φ_{2008}.

We call the first alternative whole-minus-sum integrated information Φ (Φ_{E} in [2]). The effective information φ is defined as the difference in time-delayed mutual information between the whole system and the parts. The effective information of the system beyond a certain partition $\mathcal{P}$ is

$$\varphi[X;\tau,\mathcal{P}] =: I(X_{t-\tau};X_t) - \sum_{k=1}^{r} I(M^k_{t-\tau};M^k_t)\,. \tag{15}$$

We can interpret $I(X_{t-\tau};X_t)$ as how good the system is at predicting its own future or decoding its own past (which are equivalent because mutual information is symmetric). Then, φ here can be seen as the loss in predictive power incurred by splitting the system according to $\mathcal{P}$. The details of the calculation of Φ (and the MIP) are shown in Box 1.

Φ is often regarded as a poor measure of integrated information because it can be negative, as established in known analytical results [4,5]. This is indeed conceptually awkward if Φ is seen as an absolute measure of integration between the parts of a system, though it is a reasonable property if Φ is interpreted as a "net synergy" measure [23]—quantifying to what extent the parts have shared or complementary information about the future state. That is, if $\Phi > 0$, we infer that the whole is better than the parts at predicting the future (i.e., $\Phi > 0$ is a sufficient condition), but a negative or zero Φ does not imply the opposite. Therefore, from an IIT perspective, a negative Φ can lead to the understandably confusing interpretation of a system having "negative integration", but, through a different lens (net synergy), it can be more easily interpreted as overall redundancy in the evolution of the system. See the section on integrated synergy ψ and Reference [23] for further discussion on whole-minus-sum measures.

Box 1. Calculating whole-minus-sum integrated information Φ.

$$\Phi[X;\tau] = \varphi[X;\tau,\mathcal{B}^{\mathrm{MIB}}] \tag{16a}$$

$$\mathcal{B}^{\mathrm{MIB}} = \arg_{\mathcal{B}} \min \frac{\varphi[X;\tau,\mathcal{B}]}{K(\mathcal{B})} \tag{16b}$$

$$\varphi[X;\tau,\mathcal{B}] = I(X_{t-\tau};X_t) - \sum_{k=1}^{2} I(M^k_{t-\tau};M^k_t) \tag{16c}$$

$$K(\mathcal{B}) = \min\left\{H(M^1_t), H(M^2_t)\right\} \tag{16d}$$

1. For **discrete variables**:

$$I(X_{t-\tau};X_t) = \sum_{x,x'} p(X_{t-\tau}=x, X_t=x') \log\left(\frac{p(X_{t-\tau}=x, X_t=x')}{p(X_{t-\tau}=x)\; p(X_t=x')}\right)$$

2. For **continuous, linear-Gaussian variables**:

$$I(X_{t-\tau};X_t) = \frac{1}{2}\log\left(\frac{\det\Sigma(X_t)}{\det\Sigma(X_t \mid X_{t-\tau})}\right)$$

3. For **continuous variables** with an arbitrary distribution, we must resort to the nearest-neighbour methods introduced by [24]. See reference for details.

2.2.4. Integrated Stochastic Interaction $\tilde{\Phi}$

We next consider the second alternative for Φ_{2008} for spontaneous information dynamics: integrated stochastic interaction $\tilde{\Phi}$. Also introduced in Barrett and Seth [2], this measure embodies similar concepts as Φ, with the main difference being that $\tilde{\Phi}$ utilises a definition of effective information in terms of an *increase in uncertainty* instead of in terms of a *loss of information*.

$\tilde{\Phi}$ is based on *stochastic interaction* $\tilde{\varphi}$, introduced by Ay [25]. Akin to Equation (15), we define stochastic interaction beyond partition $\mathcal{P}$ as

$$\tilde{\varphi}[X;\tau,\mathcal{P}] =: \sum_{k=1}^{r} H(M_{t-\tau}^{k} \mid M_{t}^{k}) - H(X_{t-\tau} \mid X_{t}). \tag{17}$$

Stochastic interaction quantifies to what extent uncertainty about the past is increased when the system is split in parts, compared to considering the system as a whole. The details of the calculation of $\tilde{\Phi}$ are similar to those of Φ and are described in Box 2.

The most notable advantage of $\tilde{\Phi}$ over Φ as a measure of integrated information is that $\tilde{\Phi}$ is guaranteed to be non-negative. In fact, as mentioned above, φ and $\tilde{\varphi}$ are related through the equation

$$\tilde{\varphi}[X;\tau,\mathcal{P}] = \varphi[X;\tau,\mathcal{P}] + I(M_{t}^{1};M_{t}^{2};\ldots;M_{t}^{r}), \tag{18}$$

where

$$I(M_{t}^{1};M_{t}^{2};\ldots;M_{t}^{r}) = \sum_{k=1}^{r} H(M_{t}^{k}) - H(X_{t}). \tag{19}$$

This measure is also linked to *information destruction*, as presented by Wiesner et al. [26]. The quantity $H(X_{t-\tau}|X_t)$ measures the amount of irreversibly destroyed information, since $H(X_{t-\tau}|X_t) > 0$ indicates that more than one possible past trajectory of the system converged on the same present state, making the system irreversible and indicating a loss of information about the past states. From this perspective, $\tilde{\varphi}$ can be understood as the difference between the information that is considered destroyed when the system is observed as a whole, or split into parts. Note, however, that this measure is time-symmetric when applied to a stationary system; for stationary systems, total instantaneous entropy does not change with time. Furthermore, we know that $\tilde{\Phi}$ can exceed TDMI in some cases and that it quantifies a mixture of both causal and simultaneous influences [5].

Box 2. Calculating integrated stochastic interaction $\tilde{\Phi}$.

$$\tilde{\Phi}[X;\tau] = \tilde{\varphi}[X;\tau,\mathcal{B}^{\mathrm{MIB}}] \tag{20a}$$

$$\mathcal{B}^{\mathrm{MIB}} = \arg_{\mathcal{B}} \min \frac{\tilde{\varphi}[X;\tau,\mathcal{B}]}{K(\mathcal{B})} \tag{20b}$$

$$\tilde{\varphi}[X;\tau,\mathcal{B}] = \sum_{k=1}^{2} H(M_{t-\tau}^{k} \mid M_{t}^{k}) - H(X_{t-\tau} \mid X_{t}) \tag{20c}$$

$$K(\mathcal{B}) = \min\left\{H(M_{t}^{1}), H(M_{t}^{2})\right\} \tag{20d}$$

1. For **discrete variables**:

$$H(X_{t-\tau} \mid X_{t}) = -\sum_{x,x'} p(X_{t-\tau}=x, X_{t}=x') \log\left(\frac{p(X_{t-\tau}=x, X_{t}=x')}{p(X_{t}=x')}\right)$$

2. For **continuous, linear-Gaussian variables**:

$$H(X_{t-\tau} \mid X_{t}) = \frac{1}{2}\log\det\Sigma(X_{t-\tau} \mid X_{t}) + \frac{1}{2}n\log(2\pi e)$$

3. For **continuous variables** with an arbitrary distribution, we must resort to the nearest-neighbour methods introduced by [24]. See reference for details.

2.2.5. Integrated Synergy ψ

Originally designed as a "more principled" integrated information measure [3], ψ shares some features with Φ and $\tilde{\Phi}$ but is grounded in a different branch of information theory, namely the Partial Information Decomposition (PID) framework [27]. In PID, the information that two (source) variables provide about a third (target) variable is decomposed into four non-negative terms as

$$I(X,Y;Z) = U_X(X;Z) + U_Y(Y;Z) + R(X,Y;Z) + S(X,Y;Z),$$

where U_α is the *unique information* of source α, R is the *redundancy* between both sources and S is their *synergy*.

Integrated synergy ψ is the information that the parts provide about the future of the system that is exclusively synergistic—i.e., cannot be provided by any combination of parts independently:

$$\psi[X;\tau,\mathcal{P}] =: I(X_{t-\tau};X_t) - \max_{\mathcal{P}} I_\cup(M^1_{t-\tau}, M^2_{t-\tau}, \ldots, M^r_{t-\tau}; X_t), \tag{21}$$

where

$$I_\cup(M^1_{t-\tau}, \ldots, M^r_{t-\tau}; X_t) =: \sum_{\mathcal{S}\subseteq\{M^1,\ldots,M^r\}} (-1)^{|\mathcal{S}|+1} I_\cap(\mathcal{S}^1_{t-\tau}, \ldots, \mathcal{S}^{|\mathcal{S}|}_{t-\tau}; X_t), \tag{22}$$

and $I_\cap(\mathcal{S}_1, \ldots, \mathcal{S}_{|\mathcal{S}|}; Z)$ denotes the redundant information sources $\mathcal{S}_1, \ldots, \mathcal{S}_{|\mathcal{S}|}$ have about target Z. The main problem of PID is that it is underdetermined. For example, for the case of two sources, Shannon's information theory specifies three quantities ($I(X,Y;Z)$, $I(X;Z)$, $I(Y;Z)$), whereas PID specifies four (S, R, U_X, U_Y). Therefore, a complete operational definition of ψ requires a definition of redundancy from which to construct the partial information components [27]. In this sense, the main shortcoming of ψ, inherited from PID, is that there is no agreed consensus on a definition of redundancy [23,28].

Here, we take Griffith's conceptual definition of ψ and we complement it with available definitions of redundancy (see Box 3). For the linear-Gaussian systems, we study here we use the minimum mutual information PID presented in [23,29]. Although we do not show any discrete examples here, for completeness, we provide complete formulae to calculate ψ for discrete variables using Griffith and Koch's redundancy measure [30]. Note that alternatives are available for both discrete and linear-Gaussian systems [27,31–34].

Box 3. Calculating integrated synergy ψ.

$$\psi[X;\tau,\mathcal{P}] = I(X_{t-\tau};X_t) - \max_{\mathcal{P}} I_\cup(M^1_{t-\tau}, \ldots, M^r_{t-\tau}; X_t) \tag{23}$$

1. For **discrete variables**: (following Griffith and Koch's [30] PID scheme)

$$\begin{aligned} I_\cup(M^1_{t-\tau}, \ldots, M^r_{t-\tau}; X_t) = \quad & \min_q \sum_{x,x'} q(x,x') \log\left(\frac{q(x,x')}{q(x)\,q(x')}\right) \\ \text{s.t.} \quad & q(M^i_{t-\tau}, X_t) = p(M^i_{t-\tau}, X_t) \end{aligned}$$

2. For **continuous, linear-Gaussian variables**:

$$I_\cup(M^1_{t-\tau}, \ldots, M^r_{t-\tau}; X_t) = \max_k I(M^k_{t-\tau}; X_t)$$

3. For **continuous variables** with an arbitrary distribution: unknown.

2.2.6. Decoder-Based Integrated Information Φ^*

Introduced by Oizumi et al. in Reference [5], decoder-based integrated information Φ^* takes a different approach from the previous measures. In general, Φ^* is given by

$$\Phi^*[X;\tau,\mathcal{P}] =: I(X_{t-\tau};X_t) - I^*[X;\tau,\mathcal{P}]\,, \tag{24}$$

where I^* is known as the *mismatched decoding information*, and quantifies how much information can be extracted from a variable if the receiver is using a suboptimal (or *mismatched*) decoding distribution [35,36]. This mismatched information has been used in neuroscience to quantify the contribution of neural correlations in stimulus coding [37], and can similarly be used to measure the contribution of inter-partition correlations to predictive information.

To calculate Φ^*, we formulate a restricted model q in which the correlations between partitions are ignored,

$$q(X_t|X_{t-\tau}) = \prod_i p(M_t^i|M_{t-\tau}^i), \tag{25}$$

and we calculate I^* for the case where the sender is using the full model p as an encoder and the receiver is using the restricted model q as a decoder. The details of the calculation of Φ^* and I^* are shown in Box 4. Unlike the previous measures shown in this section, Φ^* does not have an interpretable formulation in terms of simpler information-theoretic functionals like entropy and mutual information.

Calculating I^* involves a one-dimensional optimisation problem, which is straightforwardly solvable if the optimised quantity, $\tilde{I}(\beta)$, has a closed form expression [35]. For systems with continuous variables, it is in general very hard to estimate $\tilde{I}(\beta)$. However, for continuous linear-Gaussian systems and for discrete systems, $\tilde{I}(\beta)$ has an analytic closed form as a function of β if the covariance or joint probability table of the system are known, respectively. In Appendix A, we derive the formulae. (Note that the version written down in Reference [5] is incorrect, although their simulations match our results; we checked results from our derived version of the formulae versus results obtained from numerical integration, and confirmed that our derived formulae are the correct ones.) Conveniently, in both the discrete and the linear-Gaussian case, $\tilde{I}(\beta)$ is concave in β (proofs in Reference [35] and in Appendix A, respectively), which makes the optimisation significantly easier.

Box 4. Calculating decoder-based integrated information Φ^*.

$$\Phi^*[X;\tau,\mathcal{P}] = I(X_{t-\tau};X_t) - I^*[X;\tau,\mathcal{P}] \tag{26a}$$

$$I^*[X;\tau,\mathcal{P}] = \max_\beta \tilde{I}(\beta;X,\tau,\mathcal{P}) \tag{26b}$$

1. For **discrete variables**:

$$\tilde{I}(\beta;X,\tau,\mathcal{P}) = -\sum_{x'} p(X_t=x')\log\sum_x p(X_{t-\tau}=x)q(X_t=x'|X_{t-\tau}=x)^\beta + \sum_{x,x'} p(X_{t-\tau}=x,X_t=x')\log q(X_t=x'|X_{t-\tau}=x)^\beta$$

2. For **continuous, linear-Gaussian variables**: (see appendix for details)

$$\tilde{I}(\beta;X,\tau,\mathcal{P}) = \frac{1}{2}\log\left(|Q||\Sigma_x|\right) + \frac{1}{2}\operatorname{tr}\left(\Sigma_x R\right) + \beta\operatorname{tr}\left(\Pi_{x|\tilde{x}}^{-1}\Pi_{x\tilde{x}}\Pi_x^{-1}\Sigma_{\tilde{x}x}\right)$$

3. For **continuous variables** with an arbitrary distribution: unknown.

2.2.7. Geometric Integrated Information Φ_G

In Reference [4], Oizumi et al. approach the notion of dynamical complexity via yet another formalism. Their approach is based on *information geometry* [38,39]. The objects of study in information geometry are spaces of families of probability distributions, considered as differentiable (smooth) manifolds. The natural metric in information geometry is the Fisher information metric, and the KL divergence provides a natural measure of (asymmetric) distance between probability distributions. Information geometry is the application of differential geometry to the relationships and structure of probability distributions.

To quantify integrated information, Oizumi et al. [4] consider the divergence between the complete model of the system under study $p(X_{t-\tau}, X_t)$ and a *restricted model* $q(X_{t-\tau}, X_t)$ in which links between the parts of the system have been severed. This is known as the *M-projection* of the system onto the manifold of restricted models $Q = \{q : q(M_t^i|X_{t-\tau}) = q(M_t^i|M_{t-\tau}^i)\}$, and

$$\Phi_G[X;\tau,\mathcal{P}] =: \min_{q\in Q} D_{KL}\left(p(X_{t-\tau}, X_t)\|q(X_{t-\tau}, X_t)\right). \tag{27}$$

Key to this measure is that, in considering the partitioned system, it is only the connections that are cut; correlations between the parts are still allowed on the partitioned system. Although conceptually simple, Φ_G is very hard to calculate compared to all other measures we consider here (see Box 5). There is no known closed form solution for any system, and we can only find approximate numerical estimates for some systems. In particular, for discrete and linear-Gaussian variables, we can formulate Φ_G as the solution of a pure constrained multivariate optimisation problem, with the advantage that the optimisation objective is differentiable and convex [40].

Box 5. Calculating geometric integration Φ_G.

$$\Phi_G[X;\tau,\mathcal{P}] = \min_q D_{KL}(p\|q) \tag{28a}$$

$$\text{s.t. } q(M_{t+\tau}^i \mid X_t) = q(M_{t+\tau}^i \mid M_t^i). \tag{28b}$$

1. For **discrete variables**: numerically optimise the objective $D_{KL}(p\|q)$ subject to the constraints

$$\sum_{x,x'} q(X_{t-\tau} = x', X_t = x) = 1 \qquad \text{and} \qquad q(M_t^i \mid X_{t-\tau}) = q(M_t^i \mid M_{t-\tau}^i)\ \forall i.$$

2. For **continuous, linear-Gaussian variables**: numerically optimise the objective

$$\Phi_G[X;\tau,\mathcal{P}] = \min_{\Sigma(E)'} \frac{1}{2}\log\frac{|\Sigma(E)'|}{|\Sigma(E)|},$$

where $\Sigma(E) = \Sigma(X_t|X_{t-1})$, and subject to the constraints

$$\Sigma(E)' = \Sigma(E) + (A - A')\Sigma(X)(A - A')^{\mathrm{T}} \qquad \text{and}$$
$$(\Sigma(X)(A - A')\Sigma(E)'^{-1})_{ii} = 0.$$

3. For **continuous variables** with an arbitrary distribution: unknown.

2.2.8. Causal Density

Causal density (CD) is somewhat distinct from the other measures considered so far, in the sense that it is a sum of information transfers rather than a direct measure of the extent to which the whole

is greater than the parts. Nevertheless, we include it here because of its relevance and use in the dynamical complexity literature.

CD was originally defined in terms of Granger causality [14,41], but here we write it in terms of Transfer Entropy (TE), which provides a more general information-theoretic definition [42]. The conditional transfer entropy from X to Y conditioned on Z is defined as

$$\mathrm{TE}_\tau(X \to Y \mid Z) =: I(X_t; Y_{t+\tau} \mid Z_t, Y_t). \tag{29}$$

With this definition of TE, we define CD as the average pairwise conditioned TE between all variables in X,

$$\mathrm{CD}[X; \tau, \mathcal{P}] =: \frac{1}{r(r-1)} \sum_{i \neq j} \mathrm{TE}_\tau(M^i \to M^j \mid M^{[ij]}), \tag{30}$$

where $M^{[ij]}$ is the subsystem formed by all variables in X except for those in parts M^i and M^j.

In a practical sense, CD has many advantages. It has been thoroughly studied in theory [43] and applied in practice, with application domains ranging from complex systems to neuroscience [44–46]. Furthermore, there are off-the-shelf algorithms that calculate TE in discrete and continuous systems [47]. For details of the calculation of CD, see Box 6.

Causal density is a principled measure of dynamical complexity, as it vanishes for purely segregated or purely integrated systems. In a highly segregated system, there is no information transfer at all, and, in a highly integrated system, there is no transfer from one variable to another beyond the rest of the system [13]. Furthermore, CD is non-negative and upper-bounded by the total time-delayed mutual information (proof in Appendix B), therefore satisfying what other authors consider an essential requirement for a measure of integrated information [4].

Box 6. Calculating causal density CD.

$$\mathrm{CD}[X; \tau, \mathcal{P}] = \frac{1}{r(r-1)} \sum_{i \neq j} \mathrm{TE}_\tau(M^i \to M^j \mid M^{[ij]}) \tag{31}$$

1. For **discrete variables**:

$$TE_\tau(X^i \to X^j \mid X^{[ij]}) = \sum_{x,x'} p\left(X^j_{t+\tau} = x'^j, X_t = x\right) \log \left(\frac{p\left(X^j_{t+\tau} = x'^j \mid X_t = x\right)}{p\left(X^j_{t+\tau} = x'^j \mid X^j_t = x^j, X^{[ij]}_t = x^{[ij]}\right)} \right)$$

2. For **continuous, linear-Gaussian variables**:

$$TE_\tau(X^i \to X^j \mid X^{[ij]}) = \frac{1}{2} \log \left(\frac{\det \Sigma\left(X^j_{t+\tau} \mid X^j_t \oplus X^{[ij]}_t\right)}{\det \Sigma\left(X^j_{t+\tau} \mid X_t\right)} \right)$$

3. For **continuous variables** with an arbitrary distribution, we must resort to the nearest-neighbour methods introduced by [24]. See reference for details.

2.2.9. Other Measures

As already mentioned, all the measures reviewed here (besides CD) were inspired by the Φ_{2008} measure, which arose from the version of IIT laid out in Ref. [12]. The most recent version of IIT [9] is conceptually distinct, and the associated "Φ-3.0" is consequently different to the measures we consider here. The consideration of perturbation of the system, as well as all of its subsets, in both the past and

the future renders Φ-3.0 considerably more computationally expensive than other Φ measures. We do not here attempt to consider the construction of an analogue of Φ-3.0 for spontaneous information dynamics. Such an undertaking lies beyond the scope of this paper.

Recently, Tegmark [17] developed a comprehensive taxonomy of all integrated information measures that can be written as a distance between a probability distribution pertaining to the whole and one obtained as a product of probability distributions pertaining to the parts. Tegmark further identified a shortlist of candidate measures, based on a set of explicit desiderata. This shortlist overlaps with the measures we consider here, and also contains other measures which are minor variants. Of Tegmark's shortlisted measures, ϕ^{M} is equivalent to $\tilde{\Phi}$ under the system's spontaneous distribution, $\phi^{\mathrm{M}}_{kk'}$ is its state-resolved version, ϕ^{oak} is transfer entropy (which we cover here through CD), and ϕ^{npk} is not defined for continuous variables. The measures Φ_G and ψ are outside of Tegmark's classification scheme.

3. Results

All of the measures of integrated information that we have described have the potential to behave in ways which are not obvious a priori, and in a manner difficult to express analytically. While some simulations of Φ, $\tilde{\Phi}$ and CD on networks have been performed [2,13], and some analytical understanding has been achieved for Φ and $\tilde{\Phi}$ [5,17], Φ^* and Φ_G have not previously been computed on models consisting of more than two components, and ψ hasn't previously been explored at all on systems with continuous variables. In this section, we study all the measures together on small networks. We compare the behaviour of the measures, and assess the extent to which each measure reflects different forms of integration and segregation as characterised via correlations and connectivity matrices.

To recap, we consider the following six measures:

- Whole-minus-sum integrated information, Φ,
- Integrated stochastic interaction, $\tilde{\Phi}$,
- Decoder-based integrated information, Φ^*,
- Geometric integrated information, Φ_G,
- Integrated synergy, ψ,
- Causal density, CD.

We use models based on stochastic linear auto-regressive (AR) processes with Gaussian variables. These constitute appropriate models for testing the measures of integrated information. They are straightforward to parameterise and simulate, and are amenable to the formulae presented in the previous section. Mathematically, we define an AR process (of order 1) by the update equation

$$X_{t+1} = AX_t + \varepsilon_t, \tag{32}$$

where ε_t is a serially independent random sample from a zero-mean Gaussian distribution with given covariance $\Sigma(\varepsilon)$, usually referred to as the *noise* or *error term*. A particular AR process is completely specified by the coupling matrix or *network* A and the noise covariance matrix $\Sigma(\varepsilon)$. An AR process is stable, and stationary, if the spectral radius of the coupling matrix is less than 1 [48]. (The spectral radius is the largest of the absolute values of its eigenvalues.) All the example systems we consider are calibrated to be stable, so the Φ measures can be computed from their stationary statistics.

We shall consider how the measures vary with respect to: (i) the strength of connections, i.e., the magnitude of non-zero terms in the coupling matrix; (ii) the topology of the network, i.e., the arrangement of the non-zero terms in the coupling matrix; (iii) the density of connections, i.e., the density of non-zero terms in the coupling matrix; and (iv) the correlation between noise inputs to different system components, i.e., the off diagonal terms in $\Sigma(\varepsilon)$. The strength and density of connections can be thought of as reflecting, in different ways, the level of integration in the network.

The correlation between noise inputs reflects (inversely) the level of segregation, in some sense. We also, in each case, compute the control measures

- Time-delayed mutual information (TDMI), $I(X_{t-\tau}, X_t)$; and
- Average absolute correlation $\bar{\Sigma}$, defined as the average absolute value of the non-diagonal entries in the system's correlation matrix.

These simple measures quantify straightforwardly the level of interdependence between elements of the system, across time and space, respectively. TDMI captures the total information generated as the system transitions from one time-step to the next, and $\bar{\Sigma}$ is another basic measure of the level of integration.

We report the unnormalised measures minimised over even-sized bipartitions—i.e., bipartitions in which both parts have the same number of components. In doing this, we avoid conflating the effects of the choice of definition of effective information with those of the choice of partition search (see section on MIP above). See the Discussion for more details on this topic.

3.1. Key Quantities for Computing the Integrated Information Measures

To compute the integrated information measures, the stationary covariance and lagged partial covariance matrices are required. By taking the expected value of $X_t^T X_t$ with Equation (32) and given that ε_t is white noise, uncorrelated in time, one obtains that the stationary covariance matrix $\Sigma(X)$ is given by the solution to the discrete-time Lyapunov equation,

$$\Sigma(X_t) = A\,\Sigma(X_t)\,A^{\mathrm{T}} + \Sigma(\epsilon_t). \tag{33}$$

This can be easily solved numerically, for example in Matlab (R2017a, MathWorks, Natick, MA, USA) via use of the `dlyap` command. The lagged covariance can also be calculated from the parameters of the AR process as

$$\Sigma(X_{t-1}, X_t) = \langle X_t (A X_t + \varepsilon_t)^{\mathrm{T}} \rangle = \Sigma(X_t) A^{\mathrm{T}}, \tag{34}$$

and partial covariances can be obtained by applying Equation (7). Finally, we obtain the analogous quantities for the partitions by the marginalisation properties of the Gaussian distribution. Given a bipartition $X_t = \{M_t, N_t\}$, we write the covariance and lagged covariance matrices as

$$\begin{aligned} \Sigma(X_t) &= \begin{pmatrix} \Sigma(X_t)_{mm} & \Sigma(X_t)_{mn} \\ \Sigma(X_t)_{nm} & \Sigma(X_t)_{nn} \end{pmatrix}, \\ \Sigma(X_{t-1}, X_t) &= \begin{pmatrix} \Sigma(X_{t-1}, X_t)_{mm} & \Sigma(X_{t-1}, X_t)_{mn} \\ \Sigma(X_{t-1}, X_t)_{nm} & \Sigma(X_{t-1}, X_t)_{nn} \end{pmatrix}, \end{aligned} \tag{35}$$

and we simply read the partition covariance matrices as

$$\begin{aligned} \Sigma(M_t) &= \Sigma(X_t)_{mm}\,, \\ \Sigma(M_{t-1}, M_t) &= \Sigma(X_{t-1}, X_t)_{mm}\,. \end{aligned} \tag{36}$$

3.2. Two-Node Network

We begin with the simplest non-trivial AR process,

$$A = \begin{pmatrix} a & a \\ a & a \end{pmatrix}, \tag{37a}$$

$$\Sigma(\epsilon) = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}. \tag{37b}$$

Setting $a = 0.4$, we obtain the same model as depicted in Figure 3 in Reference [5]. We simulate the AR process with different levels of noise correlation c and show results for all the measures in Figure 1. Note that, as c approaches 1, the system becomes degenerate, so some matrix determinants in the formulae become zero causing some measures to diverge.

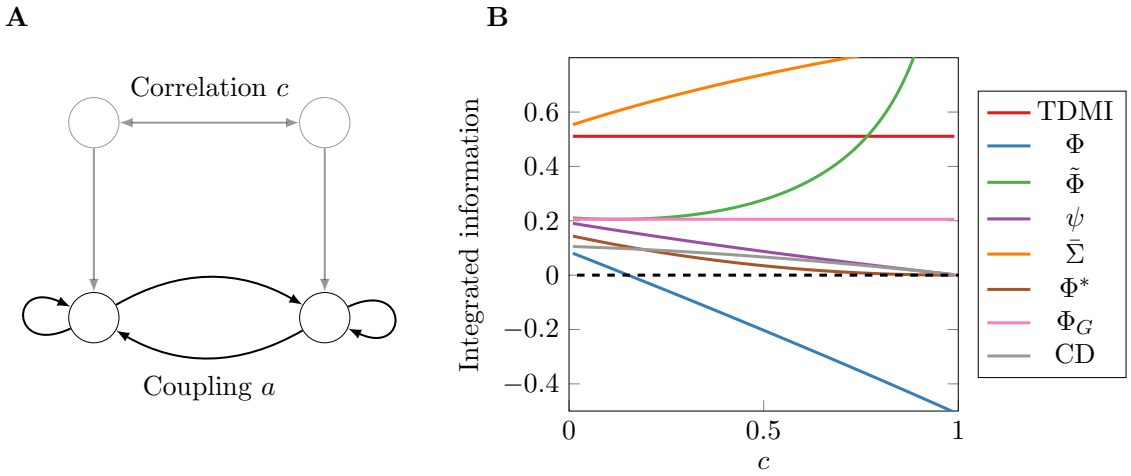

Figure 1. (**A**) graphical representation of the two-node AR process described in Equation (37). Two connected nodes with coupling strength a receive noise with correlation c, which can be thought of as coming from a common source; (**B**) all integrated information measures for different noise correlation levels c.

Inspection of Figure 1 immediately reveals a wide variability of behaviour among the measures, in both value and trend, even for this minimally simple model. A good candidate measure of (dynamical) integrated information should tend to 0 as the noise tends to becoming perfectly correlated ($c \to 1$) because, in that instance, the whole just becomes a collection of copies of the parts (we don't consider $c = 1$ because the Gaussian model becomes singular in this limit). Only the measures ψ, Φ^*, and CD achieve this. Φ_G is here unaffected by noise correlation [49], and $\tilde{\Phi}$ grows monotonically with c. Furthermore, $\tilde{\Phi}$ diverges to infinity as $c \to 1$. On the other hand, Φ also decreases monotonically but becomes negative for large enough c.

In Figure 2, we analyse the same system, but now varying both noise correlation c and coupling strength a. As per the stability condition presented above, any value of $a \geq 0.5$ makes the system's spectral radius greater than or equal to 1, so the system becomes non-stationary and variances diverge. Hence, in these plots, we evaluate all measures for values of a below the limit $a = 0.5$.

Again, the measures behave very differently. In this case, TDMI and Φ_G remain unaffected by noise correlation, and grow with increasing coupling strength as expected. In contrast, $\tilde{\Phi}$ and $\bar{\Sigma}$ increase with both a and c. Φ decreases with c but shows non-monotonic behaviour with a. Three of the measures, ψ, Φ^*, and CD, show properties consistent with capturing conjoined segregation and integration—they monotonically decrease with noise correlation and increase with coupling strength.

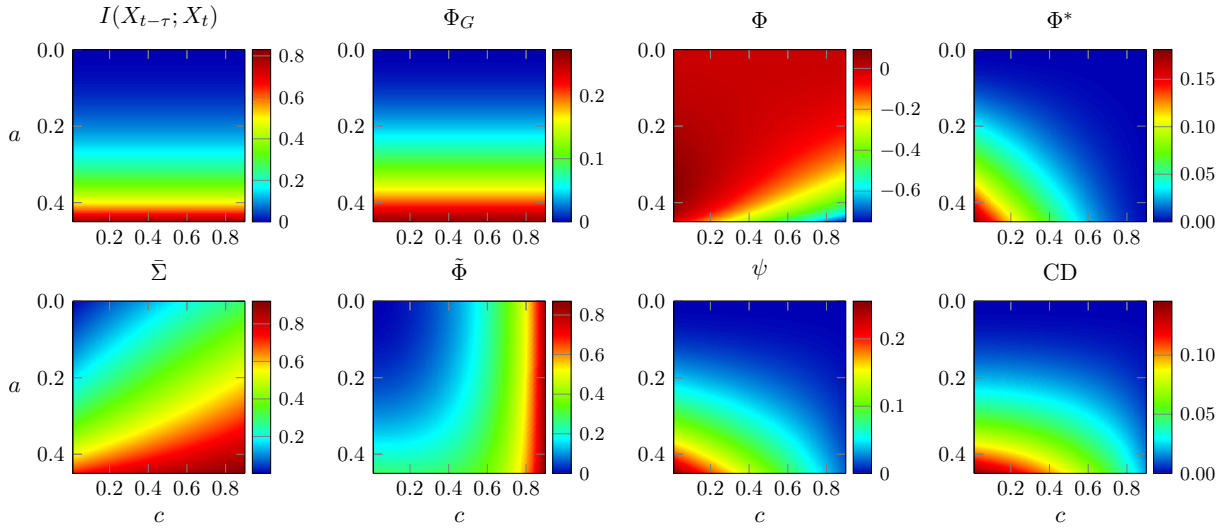

Figure 2. All integrated information measures for the two-node AR process described in Equation (37), for different coupling strengths a and noise correlation levels c. The vertical axis is inverted for visualisation purposes.

3.3. Eight-Node Networks

We now turn to networks with eight nodes, enabling examination of a richer space of dynamics and topologies.

We first analyse a weighted network optimised using a genetic algorithm to yield high Φ (Figure 2b in [2]). The noise covariance matrix has ones in the diagonal and c everywhere else, and now a is a global factor applied to all edges of the network. The (weighted) adjacency matrix is scaled such that its spectral radius is 1 when $a = 1$. Similar to the previous section, we evaluate all measures for multiple values of a and c and show the results in Figure 3.

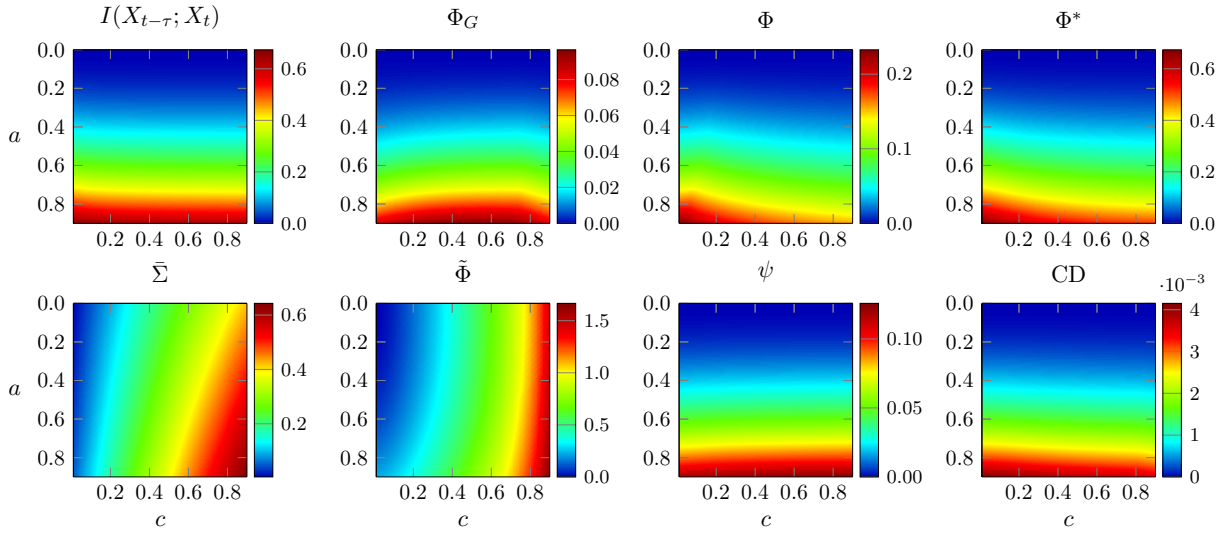

Figure 3. All integrated information measures for the Φ-optimal AR process proposed by [2], for different coupling strengths a and noise correlation levels c. Vertical axis is inverted for visualisation purposes.

Moving to a larger network mostly preserves the features highlighted above. TDMI is unaffected by c; $\tilde{\Phi}$ behaves like $\bar{\Sigma}$ and diverges for large c; and Φ^* and CD have the same trend as before, although now the decrease with c is less pronounced. Interestingly, ψ and Φ_G increase slightly with c, and Φ

does not show the instability and negative values seen in Figure 2. Overall, in this more complex network, the effect of increasing noise correlation on Φ, ψ, Φ^*, and CD is not as pronounced as in simpler networks, where these measures decrease rapidly towards zero with increasing c.

Thus far, we have studied the effect of AR dynamics on integrated information measures, keeping the topology of the network fixed and changing only global parameters. We next examine the effect of network topology, on a set of six networks:

A A fully connected network without self-loops.
B The Φ-optimal binary network presented in [2].
C The Φ-optimal weighted network presented in [2].
D A bidirectional ring network.
E A "small-world" network, formed by introducing two long-range connections to a bidirectional ring network.
F A unidirectional ring network.

In each network, the adjacency matrix has been normalised to a spectral radius of 0.9. As before, we simulate the system following Equation (32), and here set noise input correlations to zero ($c = 0$) so the noise input covariance matrix is just the identity matrix. Figure 4 shows connectivity diagrams of the networks for visual comparison, and Figure 5 shows the values of all integrated information measures evaluated on all networks.

As before, there is substantial variability in the behaviour of all measures, but some general patterns are apparent. Intriguingly, the unidirectional ring network consistently scores highest for all measures (except for $\tilde{\Phi}$ and CD), followed in most cases by the weighted Φ-optimal network [50]. On the other end of the spectrum, the fully connected network **A** consistently scores lowest, which is explained by the large correlation between its nodes as shown by $\bar{\Sigma}$.

The results here can be summarised by comparing the rank assigned to the networks by each measure (see Table 3). Inspecting this table reveals a remarkable alignment between TDMI, Φ_G, Φ^*, and ψ, especially given how much their behaviour diverges when varying a and c. Although the particular values are different, the measures largely agree on the ranking of the networks based on their integrated information. This consistency of ranking is initially encouraging with regard to empirical application.

Table 3. Networks ranked according to their value of each integrated information measure (highest value to the left). We add small-world index as a dynamics-agnostic measure of network complexity.

Measure	**Ranking**					
$I(X_t, X_{t+\tau})$	F	C	D	E	B	A
Φ_G	F	C	D	E	B	A
Φ	F	C	B	E	D	A
Φ^*	F	C	B	E	D	A
$\bar{\Sigma}$	C	B	A	E	D	F
$\tilde{\Phi}$	C	F	B	D	E	A
ψ	F	C	D	E	B	A
CD	C	F	B	D	E	A
SWI	C	E	B	A	D	F

However, the ranking is not what might be expected from topological complexity measures from network theory. If we ranked these networks by e.g., small-world index (SWI) [51–53], we expect networks **B**, **C**, and **E** to be at the top and networks **A**, **D**, and **F** to be at the bottom—very different from any of the rankings in Table 3. In fact, the Spearman correlation between the ranking by small-world index and those by TDMI, Φ_G, Φ^*, and ψ is around -0.4, leading to the conclusion that more structurally complex networks integrate *less* information. We note that these rankings are very robust to noise correlation (results not shown) for all measures except Φ. Across all simulations in this

study, the behaviour of Φ is erratic, undermining prospects for empirical application. (This behaviour is even more prevalent if Φ is optimised over all bipartitions, as opposed to over even bipartitions.)

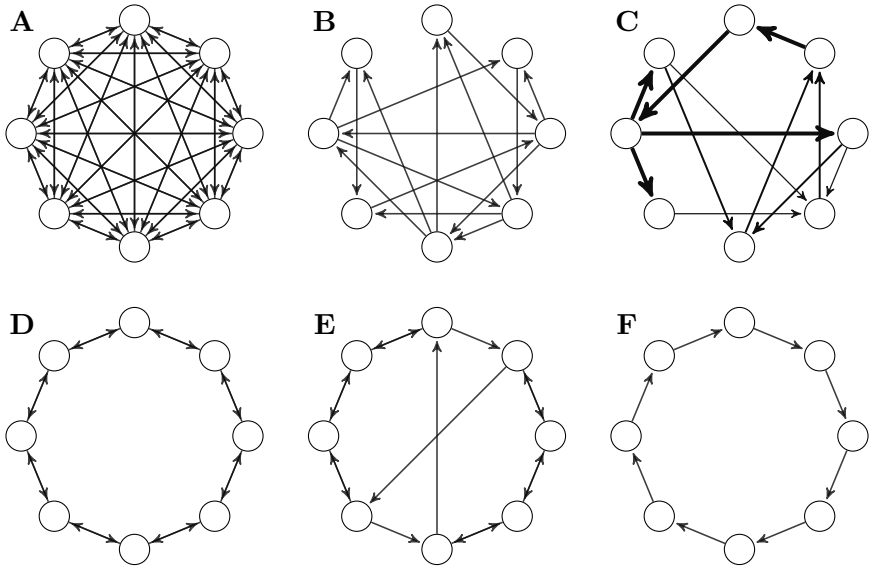

Figure 4. Networks used in the comparative analysis of integrated information measures. (**A**) fully connected network; (**B**) Φ-optimal binary network from [2]; (**C**) Φ-optimal weighted network from [2]; (**D**) bidirectional ring network; (**E**) small world network; and (**F**) is a unidirectional ring network.

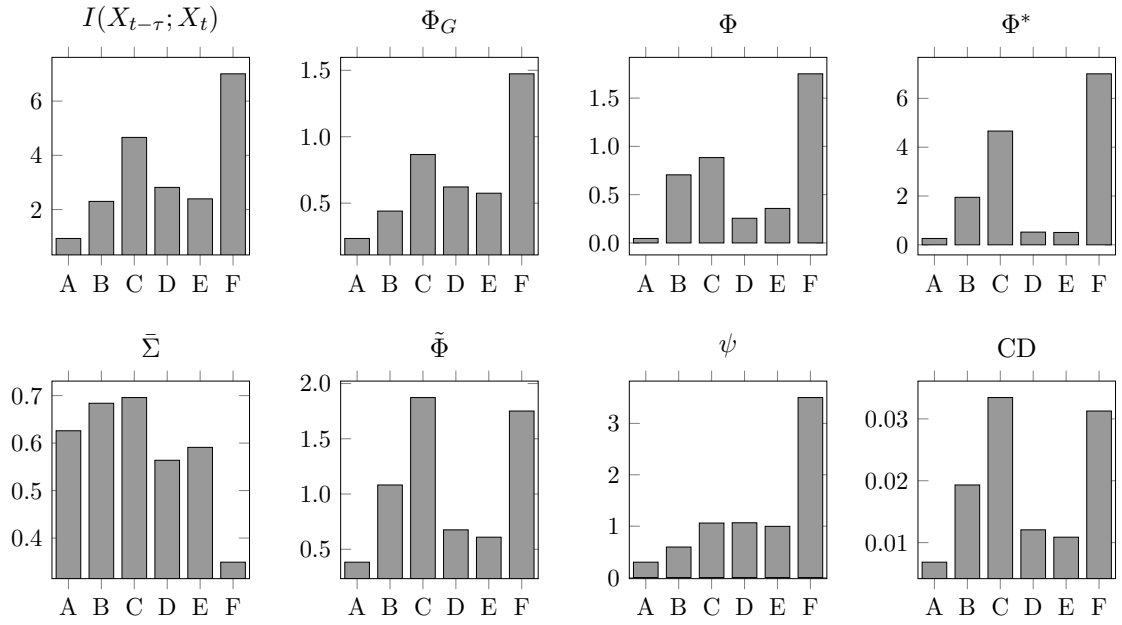

Figure 5. Integrated information measures for all networks in the suite shown in Figure 4, normalised to spectral radius 0.9 and under the influence of uncorrelated noise. The ring and weighted Φ-optimal networks score consistently at the top, while denser networks like the fully connected and the binary Φ-optimal networks are usually at the bottom. Most measures disagree on specific values but agree on the relative ranking of the networks.

3.4. Random Networks

We next perform a more general analysis of the performance of measures of integrated information, using Erdős–Rényi random networks. We consider Erdős–Rényi random networks parametrised by

two numbers: the edge density of the network ρ and the noise correlation c (defined as above), both in the $[0, 1)$ interval. To sample a network with a given ρ, we generate a matrix in which each possible edge is present with probability ρ and then remove self-loops. The stochasticity in the construction of the Erdős–Rényi network induces fluctuations on the integrated information measures, such that, for each (ρ, c), we calculate the mean and variance of each measure.

First, we generate 50 networks for each point in the (ρ, c) plane and take the mean of each integrated information measure evaluated on those 50 networks. As before, the adjacency matrices are normalised to a spectral radius of 0.9. Results are shown in Figure 6.

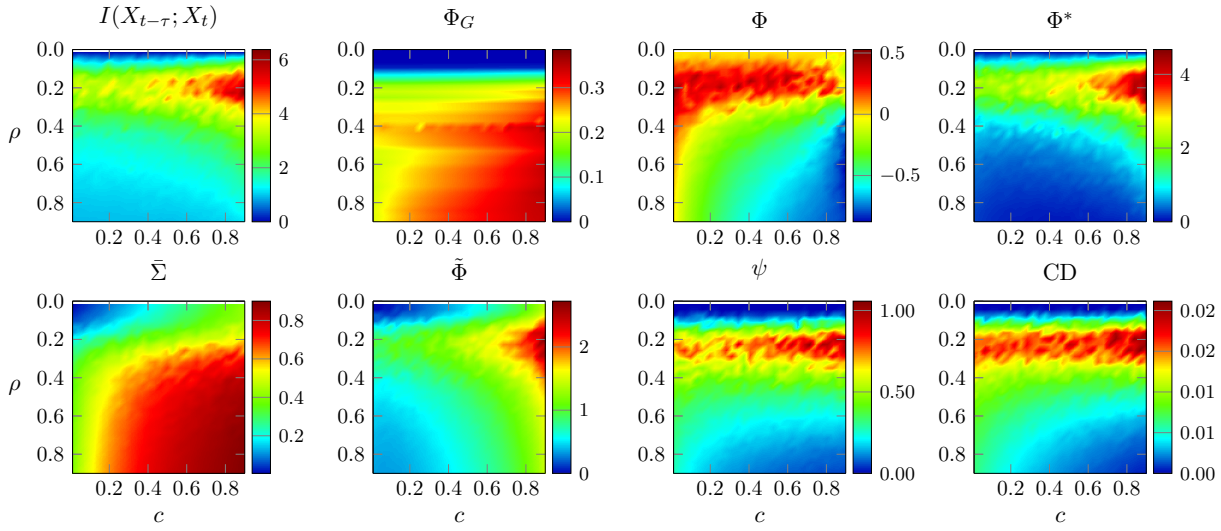

Figure 6. Average integrated information measures for Erdős–Rényi random networks with given density ρ and noise correlation c. The vertical axis is inverted for consistency with Figures 2 and 3.

Φ_G increases markedly with ρ and moderately with c, $\bar{\Sigma}$ increases sharply with both and the rest of the measures can be divided in two groups, with Φ, ψ and CD that decrease with c and TDMI, $\tilde{\Phi}$ and Φ^* that increase. Notably, all integrated information measures except Φ_G show a band of high value at an intermediate value of ρ. This demonstrates their sensitivity to the level of integration. The decrease when ρ is increased beyond a certain point is due to the weakening of the individual connections in that case (due to the fixed overall coupling strength, as quantified by spectral radius).

Secondly, in Figure 7, we plot each measure against the average correlation of each network, following the rationale that dynamical complexity should (as a necessary, but not sufficient condition) peak at an intermediate value of $\bar{\Sigma}$—i.e., it should reach its maximum value in the middle range of $\bar{\Sigma}$. To obtain this figure, we sampled a large number of Erdős–Rényi networks with random (ρ, c), and evaluated all integrated information measures, as well as their average correlation $\bar{\Sigma}$.

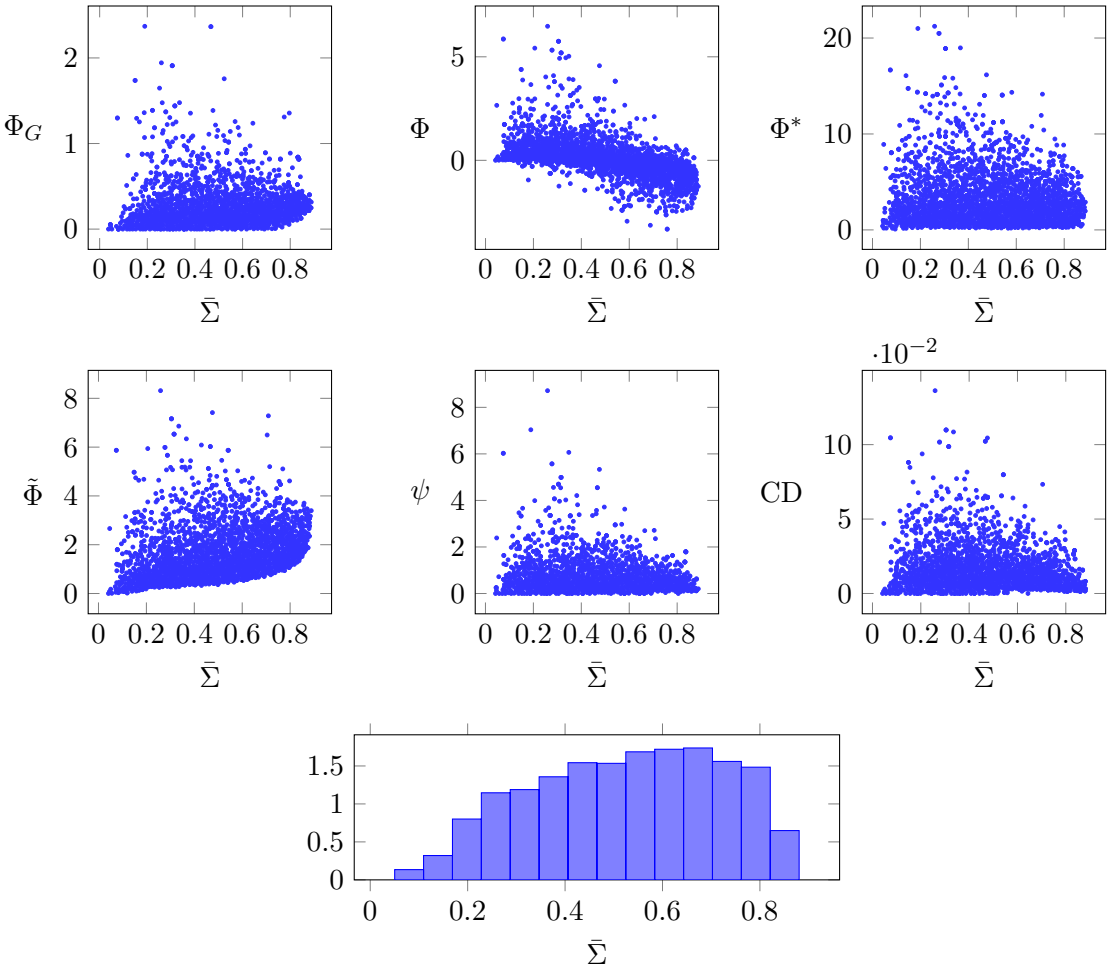

Figure 7. Integrated information measures of random Erdős–Rényi networks, plotted against the average correlation $\bar{\Sigma}$ of the same network; (bottom) normalised histogram of $\bar{\Sigma}$ for all sampled networks.

Figure 7 shows that some of the measures have this intermediate peak, in particular: Φ^*, ψ, Φ_G, and CD. Although also showing a modest intermediate peak, $\tilde{\Phi}$ has a stronger overall positive trend with $\bar{\Sigma}$, and Φ an overall negative trend. These analyses further support the notion that Φ^*, ψ, Φ_G, and CD reflect some form of dynamical complexity, although the relation between them remains unclear and not always consistent in other scenarios.

One might worry that these peaks could be due to a biased sampling of the $\bar{\Sigma}$ axis—if our sampling scheme were obtaining many more samples in, say, the $0.2 < \bar{\Sigma} < 0.4$ range, then the points with high Φ we see in that range could be explained by the fact that the high-Φ tails of the distribution are sampled better in that range than in the rest of the $\bar{\Sigma}$ axis. However, the histogram at the bottom of Figure 7 shows this is not the case—on the contrary, the samples are relatively uniformly spread along the axis. Therefore, the peaks shown by Φ^*, ψ, Φ_G, and CD are not sampling artefacts.

4. Discussion

In this study, we compared several candidate measures of integrated information in terms of their theoretical construction, and their behaviour when applied to the dynamics generated by a range of non-trivial network architectures. We found that no two measures had precisely the same basic mathematical properties (see Table 2). Empirically, we found a striking variability in the behaviour among the measures even for simple systems; see Table 4 for a summary. Three of the measures, ψ, Φ^* and CD, capture conjoined segregation and integration on small networks, when animated

with Gaussian linear AR dynamics (Figure 1). These measures decrease with increasing noise input correlation and increase with increasing coupling strength (Figure 3). Furthermore, on random networks with fixed overall coupling strength (as quantified by spectral radius), they achieve their highest scores when an intermediate number of connections are present (Figure 6). They also obtain their highest scores when the average correlation across components takes an intermediate value (Figure 7).

Table 4. Integrated information measures considered and brief summary of our results.

Measure	Summary of Results
Φ	Erratic behaviour, negative when nodes are strongly correlated.
$\tilde{\Phi}$	Mostly reflects noise input correlation, not sensitive to changes in coupling.
ψ	Reflects both segregation and integration.
Φ^*	Reflects both segregation and integration.
Φ_G	Mostly reflects changes in coupling, not sensitive to noise input correlation.
CD	Reflects both segregation and integration.

In terms of network topology, none of the measures strongly reflect complexity of the network structure in a graph theoretic sense. At fixed overall coupling strength, a simple ring structure (Figure 4) leads in most cases to the highest scores. Among the other measures: $\tilde{\Phi}$ is largely determined by the level of correlation amongst the noise inputs, and is not very sensitive to changes in coupling strength; Φ_G depends mainly on the overall coupling strength, and is not very sensitive to changes in noise input correlation; and Φ generally behaves erratically.

Considered together, our results motivate the continued development of ψ, Φ^* and CD as theoretically sound and empirically adequate measures of dynamical complexity.

4.1. Partition Selection

Integrated information is typically defined as the effective information beyond the minimum information partition [12,54]. However, when a particular measure of integrated information has been first introduced, it is often with a new operationalisation of both effective information and the minimum information partition. In this paper, we have restricted attention to comparing different choices of measure of effective information, while keeping the same partition selection scheme across all measures. Specifically, we restricted the partition search to even-sized bipartitions, which has the advantage of obviating the need for introducing a normalisation factor when comparing bipartitions with different sizes. For uneven partitions, normalisation factors are required to compensate for the fact that there is less capacity for information sharing as compared to even partitions. However, such factors are known to introduce instabilities, both under continuous parameter changes, and in terms of numerical errors [2]. Further research is needed to compare different approaches to defining the minimum information partition, or finding an approximation to it in reasonable computation time [55].

In terms of computation time, performing the most thorough search, through all partitions, as in the early formulation of Φ by Balduzzi and Tononi [12] requires time $\mathcal{O}(n^n)$. Restricting attention to bipartitions reduces this to $\mathcal{O}(2^n)$, whilst restricting to even bipartitions reduces this further to $\mathcal{O}(n^2)$. These observations highlight a trade-off between computation time and comprehensive consideration of possible partitions. Future comparisons of integrated information measures may benefit from more advanced methods for searching among a restricted set of partitions to obtain a good approximation to the minimum information partition. For example, Toker and Sommer use graph modularity, stochastic block models or spectral clustering as informed heuristics to suggest a small number of partitions likely to be close to the MIP, and then take the minimum over those. With these approximations, they are able to calculate the MIP of networks with hundreds of nodes [6,55]. Alternatively, Hidaka and Oizumi make use of the submodularity of mutual information to perform efficient optimisation and find the bipartition across which there is the least instantaneous mutual information of the system [56].

Presently, however, their method is valid only for instantaneous mutual information and is therefore not applicable to finding the bipartition that minimises any form of normalised effective information as described above in the section dedicated to the MIP.

Furthermore, each measure carries special considerations regarding partition search. For example, for ψ, taking the minimum across all partitions is equivalent to taking it across bipartitions only, thanks to the properties of $I_\cap$ [23,27,31]. Arsiwalla and Verschure [57] used $\tilde{\Phi}$ and suggested always using the atomic partition on the basis that it is fast, well-defined, and, for $\tilde{\Phi}$ specifically, it can be proven to be the partition of *maximum* information; and thus it provides a quickly computable upper bound for the measure.

4.2. Continuous Variables and the Linear Gaussian Assumption

We have compared the various integrated information measures only on systems whose states are given by continuous variables with a Gaussian distribution. This is motivated by measurement variables being best characterised as continuous in many domains of potential application. Future research should continue the comparison of these measures on a test-bed of systems with discrete variables. Moreover, non-Gaussian continuous systems should also be considered because the Gaussian approximation is not always a good fit to real data. For example, the spiking activity of populations of neurons typically exhibit exponentially distributed dynamics [58]. Systems with discrete variables are in principle straightforward to deal with, since calculating probabilities (following the most brute-force approach) amounts simply to counting occurrences of states. General continuous systems, however, are less straightforward. Estimating generic probability densities in a continuous domain is challenging, and calculating information-theoretic quantities on these is difficult [24,59]. The AR systems we have studied here are a rare exception, in the sense that their probability density can be calculated and all relevant information-theoretic quantities have an analytical expression. Nevertheless, the Gaussian assumption is common in biology, and knowing now how these measures behave on these Gaussian systems will inform further development of these measures, and motivate their application more broadly.

4.3. Empirical as Opposed to Maximum Entropy Distribution

We have considered versions of each measure that quantify information with respect to the empirical, or spontaneous, stationary distribution for the state of the system. This constitutes a significant divergence from the supposedly fundamental measures of intrinsic integrated information of IIT versions 2 and 3 [9,12]. Those measures are based on information gained about a hypothetical past moment in which the system was equally likely to be in any one of its possible states (the "maximum entropy" distribution). However, as pointed out previously [2], it is not possible to extend those measures, developed for discrete Markovian systems, to continuous systems. This is because there is no uniquely defined maximum entropy distribution for a continuous random variable (unless it has hard-bounds, i.e., a closed and bounded set of states). Hence, quantification of information with respect to the empirical distribution is the pragmatic choice for construction of an integrated information measure applicable to continuous time-series data.

The consideration of information with respect to the empirical, as opposed to maximum entropy, distribution does, however, have an effect on the concept underlying the measure of integrated information—it results in a measure not of mechanism, but of dynamics [60]. That is, what is measured is not information about what the possible mechanistic causes of the current state *could be*, but rather what the likely preceding states *actually are*, on average, statistically; see [2] for further discussion. Given the diversity of behaviour of the various integrated information measures considered here even on small networks with linear dynamics, one must remain cautious about considering them as generalisations or approximations of the proposed "fundamental" Φ measures of IIT versions 2 or 3 [9,12].

A remaining important challenge, in many practical scenarios, is the identification of stationary epochs. For a relatively long data segment, it can be unrealistic to assume that all the statistics are constant throughout. For shorter data segments, one can not be confident that the system has explored all the states that it potentially would have, given enough time.

5. Final Remarks

The further development, and empirical application of Integrated Information Theory requires a good understanding of the various potential operational measures of information integration. During the last few years, several measures have been proposed, but their behaviour in any but the simplest cases has not been extensively characterised or compared. In this study, we have reviewed several candidate measures of (dynamical/empirical) integrated information, and provided a comparative analysis on simulated data, generated by simple Gaussian dynamics applied to a range of network topologies.

Assessing the degree of dynamical complexity, integrated information, or co-existing integration and segregation exhibited by a system remains an important outstanding challenge. Progress in meeting this challenge will have implications not only for theories of consciousness, such as Integrated Information Theory, but more generally in situations where relations between local and global dynamics are of interest. The review presented here identifies promising theoretical approaches for designing adequate measures of integrated information. Furthermore, our simulations demonstrate the need for empirical investigation of such measures, since measures that share similar theoretical properties can behave in substantially different ways, even on simple systems.

Author Contributions: Conceptualization, P.A.M.M. and A.B.B.; Software and Investigation, P.A.M.M.; Writing—Original Draft Preparation, P.A.M.M. and A.B.B.; Writing—Review and Editing, P.A.M.M., A.B.B. and A.K.S.; Supervision, A.B.B. and A.K.S.

Funding: A.B.B. is funded by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/L005131/1. A.K.S. is funded by the Canadian Institute for Advanced Research, Azrieli Programme in Brain, Mind, and Consciousness. A.B.B. and A.K.S. are grateful to the Mortimer and Theresa Sackler Foundation, which supports the Sackler Centre for Consciousness Science.

Acknowledgments: The authors would like to thank Michael Schartner for advice.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A. Derivation and Concavity Proof of I^*

Appendix A.1. Derivation of I^* in Gaussian Systems

Here, we provide a closed-form expression for the mismatched decoding information in a Gaussian dynamical system. For clarity, we omit the $X, \tau, \mathcal{P}$ arguments of $\tilde{I}$ and write it as a function of β only. The formula for $\tilde{I}(\beta)$ for a stationary continuous random process is

$$\tilde{I}(\beta) = -\int \mathrm{d}x\, p(x) \log \int \mathrm{d}\tilde{x}\, p(\tilde{x}) q(x|\tilde{x})^{\beta} + \int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x}) \log q(x|\tilde{x})^{\beta}, \tag{A1}$$

where $p(x)$ is the distribution for X_t, $p(x,\tilde{x})$ is the joint distribution for $(X_t, X_{t-\tau})$, and $q(x|\tilde{x})$ is the conditional distribution for X_t given $X_{t-\tau}$ under the partitioning in question. The function $\tilde{I}(\beta)$ also depends on X_t, τ and $\mathcal{P}$, but for the sake of clarity we omit all arguments except for β, which is the parameter of interest here. When X_t is Gaussian with covariance matrix Σ_X (and mean 0 without loss of generality), we have

$$p(x) = (2\pi)^{-n/2} |\Sigma_X|^{-1/2} \exp\left[-\frac{1}{2}\psi\left(x, \Sigma_X^{-1}\right)\right], \tag{A2}$$

where we define

$$\psi(x, M) =: x^{\mathrm{T}} M x \tag{A3}$$

for a vector x and a matrix M. Furthermore,

$$q(x|\tilde{x}) \quad = \quad (2\pi)^{-n/2} |\Pi_{X|\tilde{X}}|^{-1/2} \exp\left[-\frac{1}{2} \psi\left(x - \Pi_{X\tilde{X}} \Pi_{\tilde{X}}^{-1} \tilde{x}, \Pi_{X|\tilde{X}}^{-1} \right) \right] , \tag{A4}$$

where Π_X is the block diagonal covariance matrix for X_t under the partition, $\Pi_{X\tilde{X}} =: \Sigma_q(X_t, X_{t-\tau}) = \Pi_{\tilde{X}X}^{\mathrm{T}}$ is the block diagonal auto-covariance matrix associated with the partition, and $\Pi_{X|\tilde{X}}$ is the partial covariance

$$\Pi_{X|\tilde{X}} = \Pi_X - \Pi_{X\tilde{X}} \Pi_{\tilde{X}}^{-1} \Pi_{\tilde{X}X} \, . \tag{A5}$$

We start with the integral

$$\int \mathrm{d}\tilde{x} \, p(\tilde{x}) q(x|\tilde{x})^\beta = (2\pi)^{-n\beta/2} |\Pi_{X|\tilde{X}}|^{-\beta/2} (2\pi)^{-n/2} \, |\Sigma_X|^{-1/2} \int \mathrm{d}\tilde{x} \exp(\mathcal{E}) \, , \tag{A6}$$

where

$$\mathcal{E} = \frac{1}{2} \tilde{x}^{\mathrm{T}} \Sigma_X^{-1} \tilde{x} - \frac{\beta}{2} \tilde{x}^{\mathrm{T}} \Pi_{\tilde{X}}^{-1} \Pi_{\tilde{X}X} \Pi_{X|\tilde{X}}^{-1} \Pi_{X\tilde{X}} \Pi_{\tilde{X}}^{-1} \tilde{x} + \beta x^{\mathrm{T}} \Pi_{X|\tilde{X}}^{-1} \Pi_{X\tilde{X}} \Pi_{\tilde{X}}^{-1} \tilde{x} - \frac{\beta}{2} x^{\mathrm{T}} \Pi_{X|\tilde{X}}^{-1} x \, . \tag{A7}$$

If we write

$$\mathcal{E} = -\frac{1}{2} (\tilde{x} - Bx)^{\mathrm{T}} Q (\tilde{x} - Bx) - \frac{1}{2} x^{\mathrm{T}} R_1 x \, , \tag{A8}$$

then

$$Q = \Sigma_X^{-1} + \beta \Pi_{\tilde{X}}^{-1} \Pi_{\tilde{X}X} \Pi_{X|\tilde{X}}^{-1} \Pi_{X\tilde{X}} \Pi_{\tilde{X}}^{-1} \, , \tag{A9a}$$

$$B^{\mathrm{T}} = \beta \Pi_{X|\tilde{X}}^{-1} \Pi_{X\tilde{X}} \Pi_{\tilde{X}}^{-1} Q^{-1} \, , \tag{A9b}$$

$$R_1 = \beta \Pi_{X|\tilde{X}}^{-1} - \beta^2 \Pi_{X|\tilde{X}}^{-1} \Pi_{X\tilde{X}} \Pi_{\tilde{X}}^{-1} Q^{-1} \Pi_{\tilde{X}}^{-1} \Pi_{\tilde{X}X} \Pi_{X|\tilde{X}}^{-1} \, , \tag{A9c}$$

so

$$\begin{aligned} \int \mathrm{d}\tilde{x} \exp(\mathcal{E}) &= \exp\left(-\frac{1}{2} x^{\mathrm{T}} R_1 x \right) \int \mathrm{d}y \exp\left(-\frac{1}{2} y^{\mathrm{T}} Q y \right) \\ &= \exp\left(-\frac{1}{2} x^{\mathrm{T}} R_1 x \right) (2\pi)^{n/2} |Q|^{-1/2} \, . \end{aligned} \tag{A10}$$

Hence, using Equations (A2) and (A6), we obtain the first term in Equation (A1):

$$-\int \mathrm{d}x \, p(x) \log \int \mathrm{d}\tilde{x} \, p(\tilde{x}) q(x|\tilde{x})^\beta = \frac{n\beta}{2} \log 2\pi + \frac{1}{2} \log\left(|Q| \cdot |\Sigma_X| \cdot |\Pi_{X|\tilde{X}}|^\beta \right) + \frac{1}{2} \operatorname{tr}(\Sigma_X R_1) \, . \tag{A11}$$

Now, moving on to the second term in Equation (A1),

$$\int \mathrm{d}\tilde{x} \int \mathrm{d}x \, p(x, \tilde{x}) \log q(x|\tilde{x})^\beta = -\frac{\beta n}{2} \log 2\pi - \frac{\beta}{2} \log |\Pi_{X|\tilde{X}}| - \frac{\beta}{2} I_1 \, , \tag{A12}$$

where

$$\begin{aligned} I_1 &= \int \mathrm{d}\tilde{x} \int \mathrm{d}x \; p(x,\tilde{x})\, \psi\left(x - \Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\tilde{x}, \Pi_{X|\tilde{X}}^{-1}\right) \\ &= \int \mathrm{d}x\, p(x)\, \psi\left(x, \Pi_{X|\tilde{X}}^{-1}\right) + \int \mathrm{d}\tilde{x}\, p(\tilde{x})\, \psi\left(\tilde{x}, \Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\right) \\ &\quad - 2\int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x})\, x^{\mathrm{T}}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\tilde{x} \\ &= \mathrm{tr}\left(\Pi_{X|\tilde{X}}^{-1}\Sigma_X\right) + \mathrm{tr}\left(\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\Sigma_X\right) - 2\;\mathrm{tr}\left(\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\Sigma_{\tilde{X}X}\right), \end{aligned} \tag{A13}$$

where $\Sigma_{\tilde{X}X} =: \Sigma(X_{t-\tau}, X_t)$. Thus, the second term in Equation (A1) is given by

$$\begin{aligned} \int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x}) \log q(x|\tilde{x})^{\beta} = &-\frac{\beta n}{2}\log 2\pi - \frac{\beta}{2}\log|\Pi_{X|\tilde{X}}| \\ &+ \frac{1}{2}\,\mathrm{tr}(\Sigma_X R_2) + \beta\,\mathrm{tr}\left(\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\Sigma_{\tilde{X}X}\right), \end{aligned} \tag{A14}$$

where

$$R_2 = -\beta\Pi_{X|\tilde{X}}^{-1} - \beta\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}. \tag{A15}$$

Finally, putting all the terms in Equations (A11) and (A14) together, we obtain

$$\tilde{I}(\beta) = \frac{1}{2}\log\left(|Q|\cdot|\Sigma_X|\right) + \frac{1}{2}\,\mathrm{tr}(\Sigma_X R) + \beta\;\mathrm{tr}\left(\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\Sigma_{\tilde{X}X}\right), \tag{A16}$$

where

$$Q = \Sigma_X^{-1} + \beta\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}, \tag{A17}$$

$$R = -\beta\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1} - \beta^2\Pi_{X|\tilde{X}}^{-1}\Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}Q^{-1}\Pi_{\tilde{X}}^{-1}\Pi_{\tilde{X}X}\Pi_{X|\tilde{X}}^{-1}. \tag{A18}$$

We note that this formula for $\tilde{I}(\beta)$ has been verified with numerical methods, and it is not the same as the formula reported by Oizumi et al. [5].

Appendix A.2. $\tilde{I}(\beta)$ Is Concave in β in Gaussian Systems

Throughout this proof, we will rely multiple times on the the book *Convex Optimization* by Boyd and Vandenberghe [40]. Our aim is to show that $\tilde{I}(\beta)$ is concave in β, which means it has a unique maximum and can be treated with standard convex optimisation tools. Throughout this proof, we follow Boyd and Vandenberghe's notation: a function f is said to be convex, convex downwards or concave upwards if $f(ax+by) \leq af(x) + bf(y)$, for all real non-negative a, b with $a + b = 1$.

We start with the second term in Equation (A1),

$$\int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x}) \log q(x|\tilde{x})^{\beta} = \beta \int \mathrm{d}\tilde{x} \int \mathrm{d}x\, p(x,\tilde{x}) \log q(x|\tilde{x}), \tag{A19}$$

which is linear in β. Moving to the first term, using Equation (A4) it can be rewritten as

$$\begin{aligned} -\int \mathrm{d}x\, p(x) \log\left[\int \mathrm{d}\tilde{x}\, p(\tilde{x}) q(x|\tilde{x})^{\beta}\right] = &-\int \mathrm{d}x\, p(x)\left[-\tfrac{n\beta}{2}\log 2\pi - \tfrac{\beta}{2}\log|\Pi_{X|\tilde{X}}|\right] \\ &-\int \mathrm{d}x\, p(x) \log\left[p(\tilde{x})\exp\left(-\beta f(x,\tilde{x})\right)\mathrm{d}\tilde{x}\right]. \end{aligned}$$

We see that the only nonlinear term in $\tilde{I}(\beta)$ is

$$-\int \mathrm{d}x\, p(x) \log\left[\int \mathrm{d}\tilde{x}\, p(\tilde{x}) \exp(-\beta f(x,\tilde{x}))\right], \tag{A20}$$

where

$$f(x,\tilde{x}) = \frac{1}{2}\psi\left(x - \Pi_{X\tilde{X}}\Pi_{\tilde{X}}^{-1}\tilde{x}, \Pi_{X|\tilde{X}}^{-1}\right). \tag{A21}$$

Now, we draw from two lemmas presented in [40]:

- An affine function preserves concavity, in the sense that a linear combination of convex (concave) functions is also convex (concave).
- A non-negative weighted sum preserves concavity. Since $p(x) > 0$, the outer integral in Equation (A20) preserves concavity,

With these two remarks, we know that, to prove the concavity of $\tilde{I}(\beta)$, we just need to prove the concavity of

$$-\log\left[\int \mathrm{d}\tilde{x}\, p(\tilde{x}) \exp\left(-\beta f(x,\tilde{x})\right)\right]. \tag{A22}$$

This is known as a log-sum-exp function, which, as per Section 3.1.5 of [40], is convex in β. Finally, the minus sign in the last equation flips the convexity and we conclude that $\tilde{I}(\beta)$ is concave in β.

Appendix B. Bounds on Causal Density

We now prove that causal density is upper-bounded by time-delayed mutual information, satisfying what other authors have considered a fundamental requirement for a measure of integrated information [4]. As before, we omit the arguments to CD for clarity. We begin by writing down CD in terms of mutual information:

$$\begin{aligned} \mathrm{CD} &= \frac{1}{n(n-1)} \sum_{i \neq j} \mathrm{TE}_\tau(X^i \to X^j | X^{[ij]}) \\ &= \frac{1}{n(n-1)} \sum_{i \neq j} I(X_t^i; X_{t+\tau}^j | X_t^{[i]}), \end{aligned} \tag{A23}$$

where as before $X_t^{[i]}$ represents the set of all variables in X_t except X_t^i. We will use the chain rule of mutual information [18],

$$I(X; Y, Z) = I(X; Z) + I(X; Y|Z). \tag{A24}$$

Using this chain rule and the non-negativity of mutual information, we can state that $I(X_t^i; X_{t+\tau}^j | X_t^{[i]}) \leq I(X_t; X_{t+\tau}^j)$, and therefore

$$\mathrm{CD} \leq \frac{1}{n(n-1)} \sum_{i \neq j} I(X_t; X_{t+\tau}^i). \tag{A25}$$

Also by the same chain rule, it is easy to see that $I(X_t; X_{t+\tau}^i) \leq I(X_t; X_{t+\tau})$. Then,

$$\mathrm{CD} \leq \frac{1}{n(n-1)} \sum_{i \neq j} I(X_t; X_{t+\tau}). \tag{A26}$$

Given that the sum runs across all $n(n-1)$ pairs, we arrive at our result

$$\mathrm{CD} \leq I(X_t; X_{t+\tau}). \tag{A27}$$

Appendix C. Properties of Integrated Information Measures

We prove the properties of in Table 2 of the main text. We will make use of the properties of mutual information introduced in Section 2.1, repeated here for convenience:

MI-1 $I(X;Y) = I(Y;X)$,

MI-2 $I(X;Y) \geq 0$,
MI-3 $I(f(X);g(Y)) = I(X;Y)$ for any injective functions f, g,

Appendix C.1. Whole-Minus-Sum Integrated Information Φ

Time-symmetric Follows from (MI-1).
Non-negative Proof by example. If $X_t^i = X_t^j$, we have $\Phi = (1-N)I(X_t^i; X_{t-\tau}^i) \leq 0$.
Rescaling-invariant Follows from (MI-3) when Balduzzi and Tononi's [12] normalisation factor is not used.
Bounded by TDMI Follows from (MI-2).

Appendix C.2. Integrated Stochastic Interaction $\tilde{\Phi}$

Time-symmetric Follows from $H(X_t|H_{t-\tau}) = H(X_{t-\tau}|H_t)$, which can be proved starting from the system temporal joint entropy

$$\begin{aligned} H(X_t, X_{t-\tau}) &= H(X_t|X_{t-\tau}) + H(X_{t-\tau}) \\ &= H(X_{t-\tau}, X_t) = H(X_{t-\tau}|X_t) + H(X_t), \end{aligned}$$

and using the fact that by the ergodic property $H(X_t) = H(X_{t-\tau})$. The same logic applies to all parts of the system:

Non-negative Follows from the fact that $\tilde{\Phi}$ is an M-projection (see Reference [4]).
Rescaling-invariant Follows from the non-invariance of differential entropy [18] (regardless of whether a normalisation factor is used).
Bounded by TDMI Proof by counterexample. In the two-node AR process of the main text $\tilde{\Phi} \to \infty$ as $c \to 1$, although TDMI remains finite.

Appendix C.3. Integrated Synergy ψ

Time-symmetric Proof by counterexample—for the AR system with

$$A = \begin{pmatrix} a & a \\ 0 & 0 \end{pmatrix} \quad , \quad \Sigma(\varepsilon) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

We have $\psi = \frac{1}{2}\log\left(1+a^2\right)$, while, for the time-reversed process, $\psi = \frac{1}{2}\log\left(1+a^4\right)$. Note that this proof applies only to the MMI-PID used in this paper and presented in [23].

Non-negative Follows from $I_\cup(X, Y; Z) < I(\{X, Y\}; Z)$ [27].
Rescaling-invariant Follows from (MI-3) and the fact that $I_\cap$ is also invariant (see property (**Eq**) in Section 5 of [3]).
Bounded by TDMI Follows from the non-negativity of $I_\cup$ [27].

Appendix C.4. Decoder-Based Integrated Information Φ^*

Non-negative Follows from $I^*[X; \tau, \mathcal{P}] \leq I(X_t; X_{t-\tau})$, proven in Reference [36].
Rescaling-invariant Assume that the measure is computed on a time series of rescaled data $X_t^r = X_t A$, where A is a diagonal matrix with positive real numbers. Then, its covariance is related to the covariance of the original time series as $\Sigma_X^r = \mathbb{E}\left[X_t^{r\mathrm{T}} X_t^r\right] = \mathbb{E}\left[A^{\mathrm{T}} X_t^{\mathrm{T}} X_t A\right] = A^2 \Sigma_X$. We can analogously calculate $\Pi_X, \Pi_{X\tilde{X}}, \Pi_{X|\tilde{X}}$ and easily verify that all A's cancel out, proving the invariance.
Bounded by TDMI Follows from $I^*[X; \tau, \mathcal{P}] \geq 0$, proven in Reference [36].

Appendix C.5. Geometric Integrated Information Φ_G

Time-symmetric	Follows from the symmetry in the constraints that define the manifold of restricted models Q [4].
Non-negative	Follows from the fact that Φ_G is an M-projection [4].
Rescaling-invariant	Given a Gaussian distribution p with covariance Σ_p, its M-projection in Q is another Gaussian with covariance Σ_q. Given a new distribution p' formed by rescaling some of the variables in p, the M-projection of p' is a Gaussian with covariance $A^2\Sigma_q$ with A a diagonal positive matrix (see above), which satisfies $D_{KL}(p\|q) = D_{KL}(p'\|q')$ and therefore Φ_G is invariant to rescaling.
Bounded by TDMI	TDMI can be defined as the M-projection of the full model p to a manifold of restricted models $Q^{MI} = \{q : q(X_t, X_{t-\tau}) = q(X_t)q(X_{t-\tau})\}$ [4]. The bound $\Phi_G \leq I(X_t; X_{t-\tau})$ follows from the fact that $Q^{MI} \subset Q$.

Appendix C.6. Causal Density

Time-symmetric	Follows from the non-symmetry of transfer entropy [61].
Non-negative	Re-writing CD as a sum of conditional MI terms, follows from (MI-2).
Rescaling-invariant	Follows from (MI-3).
Bounded by TDMI	Proven in S2 Appendix.

References and Notes

1. Holland, J. *Complexity: A Very Short Introduction*; Oxford University Press: Oxford, UK, 2014.
2. Barrett, A.B.; Seth, A.K. Practical measures of integrated information for time-series data. *PLoS Comput. Biol.* **2011**, *7*, e1001052. [CrossRef] [PubMed]
3. Griffith, V. A principled infotheoretic ϕ-like measure. *arXiv* **2014**, arXiv:1401.0978.
4. Oizumi, M.; Tsuchiya, N.; Amari, S.-I. A unified framework for information integration based on information geometry. *arXiv* **2015**, arXiv:1510.04455.
5. Oizumi, M.; Amari, S.-I.; Yanagawa, T.; Fujii, N.; Tsuchiya, N. Measuring integrated information from the decoding perspective. *arXiv* **2015**, arXiv:1505.04368.
6. Toker, D.; Sommer, F.T. Great than the sum: Integrated information in large brain networks. *arXiv* **2017**, arXiv:1708.02967.
7. Mediano, P.A.M.; Farah, J.C.; Shanahan, M.P. Integrated information and metastability in systems of coupled oscillators. *arXiv* **2016**, arXiv:1606.08313.
8. Tagliazucchi, E. The signatures of conscious access and its phenomenology are consistent with large-scale brain communication at criticality. *Conscious. Cogn.* **2017**, *55*, 136–147. [CrossRef] [PubMed]
9. Oizumi, M.; Albantakis, L.; Tononi, G. From the phenomenology to the mechanisms of consciousness: Integrated information theory 3.0. *PLoS Comput. Biol.* **2014**, *10*, e1003588. [CrossRef] [PubMed]
10. Tononi, G.; Sporns, O.; Edelman, G.M. A measure for brain complexity: Relating functional segregation and integration in the nervous system. *Proc. Natl. Acad. Sci. USA* **1994**, *91*, 5033–5037. [CrossRef] [PubMed]
11. Sporns, O. Complexity. *Scholarpedia* **2007**, *2*, 1623. [CrossRef]
12. Balduzzi, D.; Tononi, G. Integrated information in discrete dynamical dystems: Motivation and theoretical framework. *PLoS Comput. Biol.* **2008**, *4*, e1000091. [CrossRef] [PubMed]
13. Seth, A.K.; Barrett, A.B.; Barnett, L. Causal density and integrated information as measures of conscious level. *Philos. Trans. A* **2011**, *369*, 3748–3767. [CrossRef] [PubMed]
14. Granger, C.W.J. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **1969**, *37*, 424. [CrossRef]
15. Seth, A.K.; Izhikevich, E.; Reeke, G.N.; Edelman, G.M. Theories and measures of consciousness: An extended framework. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 10799–10804. [CrossRef] [PubMed]
16. Kanwal, M.S.; Grochow, J.A.; Ay, N. Comparing information-theoretic measures of complexity in Boltzmann machines. *Entropy* **2017**, *19*, 310. [CrossRef]
17. Tegmark, M. Improved measures of integrated information. *arXiv* **2016**, arXiv:1601.02626.
18. Cover, T.M.; Thomas, J.A. *Elements Information Theory*; Wiley: Hoboken, NJ, USA, 2006.

19. The formal derivation of the differential entropy proceeds by considering the entropy of a discrete variable with k states, and taking the $k \to \infty$ limit. The result is the differential entropy plus a divergent term that is usually dropped and is ultimately responsible for the undesirable properties of differential entropy. In the case of $I(X;Y)$ the divergent terms for the various entropies involved cancel out, restoring the useful properties of its discrete counterpart.
20. Although the origins of causal density go as back as 1969, it hasn't been until the last decade that it has found its way into neuroscience. The paper referenced in the table acts as a modern review of the properties and behaviour of causal density. This measure is somewhat distinct from the others, but is still a measure of complexity based on information dynamics between the past and current state; therefore its inclusion here will be useful.
21. Krohn, S.; Ostwald, D. Computing integrated information. *arXiv* **2016**, arXiv:1610.03627.
22. The c and e here stand respectively for cause and effect. Without an initial condition, here that the uniform distribution holds at time 0, there would be no well-defined probability distribution for these states. Further, Markovian dynamics are required for these probability distributions to be well-defined; for non-Markovian dynamics, a longer chain of initial states would have to be specified, going beyond just that at time 0.
23. Barrett, A.B. An exploration of synergistic and redundant information sharing in static and dynamical gaussian systems. *arXiv* **2014**, arXiv:1411.2832.
24. Kraskov, A.; Stögbauer, H.; Grassberger, P. Estimating mutual information. *Phys. Rev. E* **2004**, *69*, 066138. [CrossRef] [PubMed]
25. Ay, N. Information geometry on complexity and stochastic interaction. *Entropy* **2015**, *17*, 2432–2458. [CrossRef]
26. Wiesner, K.; Gu, M.; Rieper, E.; Vedral, V. Information-theoretic bound on the energy cost of stochastic simulation. *arXiv* **2011**, arXiv:1110.4217.
27. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
28. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J. Shared information—New insights and problems in decomposing information in complex systems. In *Proceedings of the European Conference on Complex Systems 2012*; Gilbert, T., Kirkilionis, M., Nicolis, G., Eds.; Springer: Berlin, Germany, 2012.
29. Barrett's derivation of the MMI-PID, which follows Williams and Beer and Griffith and Koch's procedure, gives this formula when the target is univariate. We generalise the formula here to the case of multivariate target in order to render ψ computable for Gaussians. This formula leads to synergy being the extra information contributed by the weaker source given the stronger source was previously known.
30. Griffith, V.; Koch, C. Quantifying synergistic mutual information. *arXiv* **2012**, arXiv:1205.4265.
31. Rosas, F.; Ntranos, V.; Ellison, C.; Pollin, S.; Verhelst, M. Understanding interdependency through complex information sharing. *Entropy* **2016**, *18*, 38. [CrossRef]
32. Ince, R.A.A. Measuring multivariate redundant information with pointwise common change in surprisal. *Entropy* **2017**, *19*, 318. [CrossRef]
33. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183. [CrossRef]
34. Kay, J.W.; Ince, R.A.A. Exact partial information decompositions for Gaussian systems based on dependency constraints. *arXiv* **2018**, arXiv:1803.02030.
35. Latham, P.E.; Nirenberg, S. Synergy, redundancy, and independence in population codes, revisited. *J. Neurosci.* **2005**, *25*, 5195–206. [CrossRef] [PubMed]
36. Merhav, N.; Kaplan, G.; Lapidoth, A.; Shitz, S.S. On information rates for mismatched decoders. *IEEE Trans. Inf. Theory* **1994**, *40*, 1953–1967. [CrossRef]
37. Oizumi, M.; Ishii, T.; Ishibashi, K.; Hosoya, T.; Okada, M. Mismatched decoding in the brain. *J. Neurosci.* **2010**, *30*, 4815–4826. [CrossRef]
38. Amari, S.-I.; Nagaoka, H. *Methods of Information Geometry*; American Mathematical Society: Providence, RI, USA, 2000.
39. Amari, S.-I. Information geometry in optimization, machine learning and statistical inference. *Front. Electr. Electron. Eng. China* **2010**, *5*, 241–260. [CrossRef]
40. Boyd, S.S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.

41. Seth, A.K. Causal connectivity of evolved neural networks during behavior. *Netw. Comput. Neural Syst.* **2005**, *16*, 35–54. [CrossRef]
42. Barnett, L.; Barrett, A.B.; Seth, A.K. Granger causality and transfer entropy are equivalent for Gaussian variables. *Phys. Rev. Lett.* **2009**, *103*, 238701. [CrossRef] [PubMed]
43. Barnett, L.; Seth, A.K. Behaviour of granger causality under filtering: Theoretical invariance and practical application. *J. Neurosci. Methods* **2011**, *201*, 404–419. [CrossRef] [PubMed]
44. Lindner, M.; Vicente, R.; Priesemann, V.; Wibral, M. TRENTOOL: A matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC Neurosci.* **2011**, *12*, 119. [CrossRef] [PubMed]
45. Lizier, J.T.; Heinzle, J.; Horstmann, A.; Haynes, J.-D.; Prokopenko, M. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fMRI connectivity. *J. Comput. Neurosci.* **2010**, *30*, 85–107. [CrossRef] [PubMed]
46. Mediano, P.A.M.; Shanahan, M.P. Balanced information storage and transfer in modular spiking neural networks. *arXiv* **2017**, arXiv:1708.04392.
47. Barnett, L.; Seth, A.K. The MVGC multivariate granger causality toolbox: A new approach to granger-causal inference. *J. Neurosci. Methods* **2014**, *223*, 50–68. [CrossRef] [PubMed]
48. Lütkepohl, H. *New Introduction to Multiple Time Series Analysis*; Springer: New York, NY, USA, 2005.
49. According to an anonymous reviewer, Φ_G does decrease with noise correlation in discrete systems, although in this article we focus exclusively in Gaussian systems.
50. Note that in Figure 5 the Φ-optimal networks **B** and **C** score much less than simpler network **F**. This is because all networks have been scaled to a spectral radius of 0.9—when the networks are normalised to a spectral radius of 0.5, as in the original paper, then **B** and **C** are, as expected, the networks with highest Φ.
51. Humphries, M.D.; Gurney, K. Network 'small-world-ness:' A quantitative method for determining canonical network equivalence. *PLoS ONE* **2008**, *3*, e0002051. [CrossRef] [PubMed]
52. Yin, H.; Benson, A.R.; Leskovec, J. Higher-order clustering in networks. *arXiv* **2017**, arXiv:1704.03913.
53. The small-world index of a network is defined as the ratio between its clustering coefficient and its mean minimum path length, normalised by the expected value of these measures on a random network of the same density. Since the networks we consider are small and sparse, we use the 4th order cliques (instead of triangles, which are 3rd order cliques) to calculate the clustering coefficient.
54. Tononi, G.; Sporns, O. Measuring information integration. *BMC Neurosci.* **2003**, *4*, 31. [CrossRef]
55. Toker, D.; Sommer, F. Moving past the minimum information partition: How to quickly and accurately calculate integrated information. *arXiv* **2016**, arXiv:1605.01096.
56. Hidaka, S.; Oizumi, M. Fast and exact search for the partition with minimal information loss. *arXiv* **2017**, arXiv:1708.01444.
57. Arsiwalla, X.D.; Verschure, P.F.M.J. Integrated information for large complex networks. In Proceedings of the 2013 International Joint Conference on Neural Networks (IJCNN), Dallas, TX, USA, 4–9 August 2013; pp. 1–7.
58. Dayan, P.; Abbott, L.F. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*; MIT Press: Cambridge, MA, USA, 2001.
59. Wang, Q.; Kulkarni, S.R.; Verdu, S. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Trans. Inf. Theory* **2009**, *55*, 2392–2405. [CrossRef]
60. Barrett, A.B.; Barnett, L. Granger causality is designed to measure effect, not mechanism. *Front. Neuroinform.* **2013**, *7*, 6. [CrossRef]
61. Wibral, M.; Vicente, R.; Lizier, J.T. (Eds.) *Directed Information Measures in Neuroscience*; Understanding Complex Systems; Springer: Berlin/Heidelberg, Germany, 2014.